

Contents

| | |
|--|-----------|
| Preface | xi |
| 1 Introduction | 1 |
| 1.1 What is machine learning? | 1 |
| 1.2 Supervised learning | 1 |
| 1.2.1 Classification | 2 |
| 1.2.2 Regression | 8 |
| 1.2.3 Overfitting and generalization | 12 |
| 1.2.4 No free lunch theorem | 13 |
| 1.3 Unsupervised learning | 13 |
| 1.3.1 Clustering | 14 |
| 1.3.2 Discovering latent “factors of variation” | 14 |
| 1.3.3 Self-supervised learning | 15 |
| 1.3.4 Evaluating unsupervised learning | 15 |
| 1.4 Reinforcement learning | 16 |
| 1.5 Data | 18 |
| 1.5.1 Some common image datasets | 18 |
| 1.5.2 Some common text datasets | 21 |
| 1.5.3 Preprocessing discrete input data | 21 |
| 1.5.4 Preprocessing text data | 22 |
| 1.5.5 Handling missing data | 25 |
| 1.6 Discussion | 25 |
| 1.6.1 The relationship between ML and other fields | 26 |
| 1.6.2 Structure of the book | 26 |
| 1.6.3 Caveats | 26 |
| I Foundations | 29 |
| 2 Probability: univariate models | 31 |
| 2.1 Introduction | 31 |
| 2.1.1 What is probability? | 31 |

| | | | |
|----------|---|-----------|--|
| 2.1.2 | Types of uncertainty | 31 | |
| 2.1.3 | Probability as an extension of logic | 32 | |
| 2.2 | Random variables | 33 | |
| 2.2.1 | Discrete random variables | 33 | |
| 2.2.2 | Continuous random variables | 34 | |
| 2.2.3 | Sets of related random variables | 36 | |
| 2.2.4 | Independence and conditional independence | 37 | |
| 2.2.5 | Moments of a distribution | 38 | |
| 2.3 | Bayes' rule | 41 | |
| 2.3.1 | Example: Testing for COVID-19 | 42 | |
| 2.3.2 | Example: The Monty Hall problem | 43 | |
| 2.3.3 | Inverse problems * | 45 | |
| 2.4 | Bernoulli and binomial distributions | 45 | |
| 2.4.1 | Definition | 46 | |
| 2.4.2 | Sigmoid (logistic) function | 46 | |
| 2.4.3 | Binary logistic regression | 48 | |
| 2.5 | Categorical and multinomial distributions | 49 | |
| 2.5.1 | Definition | 49 | |
| 2.5.2 | Softmax function | 50 | |
| 2.5.3 | Multiclass logistic regression | 51 | |
| 2.5.4 | Log-sum-exp trick | 52 | |
| 2.6 | Univariate Gaussian (normal) distribution | 53 | |
| 2.6.1 | Cumulative distribution function | 53 | |
| 2.6.2 | Probability density function | 54 | |
| 2.6.3 | Regression | 55 | |
| 2.6.4 | Why is the Gaussian distribution so widely used? | 56 | |
| 2.6.5 | Dirac delta function as a limiting case | 57 | |
| 2.7 | Some other common univariate distributions * | 57 | |
| 2.7.1 | Student t distribution | 57 | |
| 2.7.2 | Cauchy distribution | 59 | |
| 2.7.3 | Laplace distribution | 59 | |
| 2.7.4 | Beta distribution | 60 | |
| 2.7.5 | Gamma distribution | 60 | |
| 2.8 | Transformations of random variables * | 62 | |
| 2.8.1 | Discrete case | 62 | |
| 2.8.2 | Continuous case | 62 | |
| 2.8.3 | Invertible transformations (bijections) | 62 | |
| 2.8.4 | Moments of a linear transformation | 64 | |
| 2.8.5 | The convolution theorem | 65 | |
| 2.8.6 | Central limit theorem | 67 | |
| 2.8.7 | Monte Carlo approximation | 68 | |
| 2.9 | Exercises | 69 | |
| 3 | Probability: multivariate models | 73 | |
| 3.1 | Joint distributions for multiple random variables | 73 | |

| | | | |
|----------|--|-----------|--|
| 3.1.1 | Covariance | 73 | |
| 3.1.2 | Correlation | 74 | |
| 3.1.3 | Uncorrelated does not imply independent | 75 | |
| 3.1.4 | Correlation does not imply causation | 75 | |
| 3.1.5 | Simpsons' paradox | 76 | |
| 3.2 | The multivariate Gaussian (normal) distribution | 77 | |
| 3.2.1 | Definition | 77 | |
| 3.2.2 | Mahalanobis distance | 79 | |
| 3.2.3 | Marginals and conditionals of an MVN * | 80 | |
| 3.2.4 | Example: Imputing missing values * | 81 | |
| 3.3 | Linear Gaussian systems * | 81 | |
| 3.3.1 | Example: inferring a latent vector from a noisy sensor | 83 | |
| 3.3.2 | Example: inferring a latent vector from multiple noisy sensors | 84 | |
| 3.4 | Mixture models | 84 | |
| 3.4.1 | Gaussian mixture models | 85 | |
| 3.4.2 | Mixtures of Bernoullis | 89 | |
| 3.4.3 | Gaussian scale mixtures * | 89 | |
| 3.5 | Probabilistic graphical models * | 90 | |
| 3.5.1 | Representation | 90 | |
| 3.5.2 | Inference | 93 | |
| 3.5.3 | Learning | 95 | |
| 3.6 | Exercises | 97 | |
| 4 | Statistics | 99 | |
| 4.1 | Introduction | 99 | |
| 4.2 | Maximum likelihood estimation (MLE) | 99 | |
| 4.2.1 | Definition | 99 | |
| 4.2.2 | Justification for MLE | 100 | |
| 4.2.3 | Example: MLE for the Bernoulli distribution | 102 | |
| 4.2.4 | Example: MLE for the categorical distribution | 103 | |
| 4.2.5 | Example: MLE for the univariate Gaussian | 103 | |
| 4.2.6 | Example: MLE for the multivariate Gaussian | 104 | |
| 4.2.7 | Example: MLE for linear regression | 106 | |
| 4.3 | Empirical risk minimization (ERM) | 107 | |
| 4.3.1 | Example: minimizing the misclassification rate | 107 | |
| 4.3.2 | Surrogate loss | 108 | |
| 4.4 | Other estimation methods * | 108 | |
| 4.4.1 | The method of moments | 108 | |
| 4.4.2 | Online (recursive) estimation | 110 | |
| 4.5 | Regularization | 112 | |
| 4.5.1 | Example: MAP estimation for the Bernoulli distribution | 113 | |
| 4.5.2 | Example: MAP estimation for the multivariate Gaussian * | 114 | |
| 4.5.3 | Example: weight decay | 115 | |
| 4.5.4 | Picking the regularizer using a validation set | 117 | |
| 4.5.5 | Cross-validation | 117 | |

| | | | |
|----------|---|------------|--|
| 4.5.6 | Early stopping | 119 | |
| 4.5.7 | Using more data | 120 | |
| 4.6 | Bayesian statistics * | 121 | |
| 4.6.1 | Conjugate priors | 121 | |
| 4.6.2 | The beta-binomial model | 122 | |
| 4.6.3 | The Dirichlet-multinomial model | 129 | |
| 4.6.4 | The Gaussian-Gaussian model | 132 | |
| 4.6.5 | Beyond conjugate priors | 135 | |
| 4.6.6 | Credible intervals | 137 | |
| 4.6.7 | Bayesian machine learning | 138 | |
| 4.6.8 | Computational issues | 142 | |
| 4.7 | Frequentist statistics * | 145 | |
| 4.7.1 | Sampling distributions | 145 | |
| 4.7.2 | Gaussian approximation of the sampling distribution of the MLE | 146 | |
| 4.7.3 | Bootstrap approximation of the sampling distribution of any estimator | 147 | |
| 4.7.4 | Confidence intervals | 148 | |
| 4.7.5 | Caution: Confidence intervals are not credible | 149 | |
| 4.7.6 | The bias-variance tradeoff | 150 | |
| 4.8 | Exercises | 155 | |
| 5 | Decision theory | 159 | |
| 5.1 | Bayesian decision theory | 159 | |
| 5.1.1 | Basics | 159 | |
| 5.1.2 | Classification problems | 161 | |
| 5.1.3 | ROC curves | 162 | |
| 5.1.4 | Precision-recall curves | 165 | |
| 5.1.5 | Regression problems | 168 | |
| 5.1.6 | Probabilistic prediction problems | 169 | |
| 5.2 | A/B testing * | 171 | |
| 5.2.1 | A Bayesian approach | 172 | |
| 5.2.2 | Example | 174 | |
| 5.3 | Bandit problems * | 176 | |
| 5.3.1 | Contextual bandits | 176 | |
| 5.3.2 | Markov decision processes | 177 | |
| 5.3.3 | Exploration-exploitation tradeoff | 178 | |
| 5.3.4 | Optimal solution | 178 | |
| 5.3.5 | Regret | 180 | |
| 5.3.6 | Upper confidence bounds (UCB) | 181 | |
| 5.3.7 | Thompson sampling | 182 | |
| 5.3.8 | Simple heuristics | 183 | |
| 5.4 | Bayesian hypothesis testing | 184 | |
| 5.4.1 | Example: Testing if a coin is fair | 185 | |
| 5.4.2 | Bayesian model selection | 185 | |
| 5.4.3 | Occam's razor | 186 | |
| 5.4.4 | Connection between cross validation and marginal likelihood | 189 | |

| | | | |
|----------|--|------------|--|
| 5.4.5 | Information criteria * | 190 | |
| 5.5 | Frequentist decision theory | 192 | |
| 5.5.1 | Computing the risk of an estimator | 192 | |
| 5.5.2 | Consistent estimators | 195 | |
| 5.5.3 | Admissible estimators | 195 | |
| 5.6 | Empirical risk minimization | 196 | |
| 5.6.1 | Empirical risk | 196 | |
| 5.6.2 | Structural risk | 198 | |
| 5.6.3 | Cross-validation | 199 | |
| 5.6.4 | Statistical learning theory * | 200 | |
| 5.7 | Frequentist hypothesis testing * | 201 | |
| 5.7.1 | Likelihood ratio test | 201 | |
| 5.7.2 | Null hypothesis significance testing (NHST) | 202 | |
| 5.7.3 | p-values | 203 | |
| 5.7.4 | p-values considered harmful | 203 | |
| 5.7.5 | Why isn't everyone a Bayesian? | 206 | |
| 5.8 | Exercises | 207 | |
| 6 | Information theory | 209 | |
| 6.1 | Entropy | 209 | |
| 6.1.1 | Entropy for discrete random variables | 209 | |
| 6.1.2 | Cross entropy | 211 | |
| 6.1.3 | Joint entropy | 211 | |
| 6.1.4 | Conditional entropy | 212 | |
| 6.1.5 | Perplexity | 213 | |
| 6.1.6 | Differential entropy for continuous random variables * | 214 | |
| 6.2 | Relative entropy (KL divergence) * | 215 | |
| 6.2.1 | Definition | 215 | |
| 6.2.2 | Interpretation | 216 | |
| 6.2.3 | Example: KL divergence between two Gaussians | 216 | |
| 6.2.4 | Non-negativity of KL | 216 | |
| 6.2.5 | KL divergence and MLE | 217 | |
| 6.2.6 | Forward vs reverse KL | 218 | |
| 6.3 | Mutual information * | 219 | |
| 6.3.1 | Definition | 219 | |
| 6.3.2 | Interpretation | 219 | |
| 6.3.3 | Example | 220 | |
| 6.3.4 | Conditional mutual information | 221 | |
| 6.3.5 | MI as a “generalized correlation coefficient” | 221 | |
| 6.3.6 | Normalized mutual information | 222 | |
| 6.3.7 | Maximal information coefficient | 223 | |
| 6.3.8 | Data processing inequality | 225 | |
| 6.3.9 | Sufficient Statistics | 226 | |
| 6.3.10 | Fano's inequality * | 226 | |
| 6.4 | Exercises | 227 | |

| | | |
|----------|--|------------|
| 7 | Linear algebra | 231 |
| 7.1 | Introduction | 231 |
| 7.1.1 | Notation | 231 |
| 7.1.2 | Vector spaces | 234 |
| 7.1.3 | Norms of a vector and matrix | 236 |
| 7.1.4 | Properties of a matrix | 238 |
| 7.1.5 | Special types of matrices | 240 |
| 7.2 | Matrix multiplication | 244 |
| 7.2.1 | Vector-Vector Products | 244 |
| 7.2.2 | Matrix-Vector Products | 244 |
| 7.2.3 | Matrix-Matrix Products | 245 |
| 7.2.4 | Application: manipulating data matrices | 247 |
| 7.2.5 | Kronecker products * | 249 |
| 7.2.6 | Einstein summation * | 250 |
| 7.3 | Matrix inversion | 251 |
| 7.3.1 | The inverse of a square matrix | 251 |
| 7.3.2 | Schur complements * | 251 |
| 7.3.3 | The matrix inversion lemma * | 253 |
| 7.3.4 | Matrix determinant lemma * | 253 |
| 7.4 | Eigenvalue decomposition (EVD) | 254 |
| 7.4.1 | Basics | 254 |
| 7.4.2 | Diagonalization | 255 |
| 7.4.3 | Eigenvalues and eigenvectors of symmetric matrices | 255 |
| 7.4.4 | Geometry of quadratic forms | 256 |
| 7.4.5 | Standardizing and whitening data | 256 |
| 7.4.6 | Power method | 258 |
| 7.4.7 | Deflation | 259 |
| 7.4.8 | Eigenvectors optimize quadratic forms | 259 |
| 7.5 | Singular value decomposition (SVD) | 259 |
| 7.5.1 | Basics | 259 |
| 7.5.2 | Connection between SVD and EVD | 260 |
| 7.5.3 | Pseudo inverse | 261 |
| 7.5.4 | SVD and the range and null space of a matrix * | 262 |
| 7.5.5 | Truncated SVD | 263 |
| 7.6 | Other matrix decompositions * | 264 |
| 7.6.1 | LU factorization | 264 |
| 7.6.2 | QR decomposition | 264 |
| 7.6.3 | Cholesky decomposition | 265 |
| 7.7 | Solving systems of linear equations * | 266 |
| 7.7.1 | Solving square systems | 267 |
| 7.7.2 | Solving underconstrained systems (least norm estimation) | 267 |
| 7.7.3 | Solving overconstrained systems (least squares estimation) | 268 |
| 7.8 | Matrix calculus | 269 |
| 7.8.1 | Derivatives | 269 |
| 7.8.2 | Gradients | 270 |

| | | |
|----------|--|------------|
| 7.8.3 | Directional derivative | 271 |
| 7.8.4 | Total derivative * | 271 |
| 7.8.5 | Jacobian | 271 |
| 7.8.6 | Hessian | 272 |
| 7.8.7 | Gradients of commonly used functions | 272 |
| 7.8.8 | Functional derivative notation * | 274 |
| 7.9 | Exercises | 277 |
| 8 | Optimization | 279 |
| 8.1 | Introduction | 279 |
| 8.1.1 | Local vs global optimization | 279 |
| 8.1.2 | Constrained vs unconstrained optimization | 281 |
| 8.1.3 | Convex vs nonconvex optimization | 281 |
| 8.1.4 | Smooth vs nonsmooth optimization | 285 |
| 8.2 | First-order methods | 286 |
| 8.2.1 | Descent direction | 287 |
| 8.2.2 | Step size (learning rate) | 288 |
| 8.2.3 | Convergence rates | 290 |
| 8.2.4 | Momentum methods | 291 |
| 8.3 | Second-order methods | 292 |
| 8.3.1 | Newton's method | 293 |
| 8.3.2 | BFGS and other quasi-Newton methods | 294 |
| 8.3.3 | Trust region methods | 295 |
| 8.3.4 | Natural gradient descent * | 296 |
| 8.4 | Stochastic gradient descent | 299 |
| 8.4.1 | Application to finite sum problems | 299 |
| 8.4.2 | Example: SGD for fitting linear regression | 300 |
| 8.4.3 | Choosing the step size (learning rate) | 301 |
| 8.4.4 | Iterate averaging | 303 |
| 8.4.5 | Variance reduction * | 303 |
| 8.4.6 | Preconditioned SGD | 305 |
| 8.5 | Constrained optimization | 307 |
| 8.5.1 | Lagrange multipliers | 308 |
| 8.5.2 | The KKT conditions | 309 |
| 8.5.3 | Linear programming | 311 |
| 8.5.4 | Quadratic programming | 312 |
| 8.5.5 | Mixed integer linear programming * | 313 |
| 8.6 | Proximal gradient method * | 313 |
| 8.6.1 | Projected gradient descent | 314 |
| 8.6.2 | Proximal operator for ℓ_1 -norm regularizer | 315 |
| 8.6.3 | Proximal operator for quantization | 316 |
| 8.7 | Bound optimization * | 317 |
| 8.7.1 | The general algorithm | 317 |
| 8.7.2 | The EM algorithm | 318 |
| 8.7.3 | Example: EM for a GMM | 321 |

| | | |
|-----------|--|------------|
| 8.7.4 | Example: EM for an MVN with missing data | 325 |
| 8.8 | Blackbox and derivative free optimization | 328 |
| 8.8.1 | Grid search and random search | 328 |
| 8.8.2 | Simulated annealing * | 328 |
| 8.8.3 | Model-based blackbox optimization * | 329 |
| 8.9 | Exercises | 330 |
| II | Linear models | 331 |
| 9 | Linear discriminant analysis | 333 |
| 9.1 | Introduction | 333 |
| 9.2 | Gaussian discriminant analysis | 333 |
| 9.2.1 | Quadratic decision boundaries | 334 |
| 9.2.2 | Linear decision boundaries | 335 |
| 9.2.3 | The connection between LDA and logistic regression | 335 |
| 9.2.4 | Model fitting | 336 |
| 9.2.5 | Nearest centroid classifier | 338 |
| 9.2.6 | Fisher's linear discriminant analysis * | 338 |
| 9.3 | Naive Bayes classifiers | 342 |
| 9.3.1 | Example models | 342 |
| 9.3.2 | Model fitting | 343 |
| 9.3.3 | Bayesian naive Bayes | 344 |
| 9.3.4 | The connection between naive Bayes and logistic regression | 345 |
| 9.4 | Generative vs discriminative classifiers | 346 |
| 9.4.1 | Advantages of discriminative classifiers | 346 |
| 9.4.2 | Advantages of generative classifiers | 347 |
| 9.4.3 | Handling missing features | 347 |
| 9.5 | Exercises | 348 |
| 10 | Logistic regression | 349 |
| 10.1 | Introduction | 349 |
| 10.2 | Binary logistic regression | 349 |
| 10.2.1 | Linear classifiers | 349 |
| 10.2.2 | Nonlinear classifiers | 350 |
| 10.2.3 | Maximum likelihood estimation | 352 |
| 10.2.4 | Stochastic gradient descent | 355 |
| 10.2.5 | Perceptron algorithm | 355 |
| 10.2.6 | Iteratively reweighted least squares | 356 |
| 10.2.7 | MAP estimation | 357 |
| 10.2.8 | Standardization | 359 |
| 10.3 | Multinomial logistic regression | 360 |
| 10.3.1 | Linear and nonlinear classifiers | 360 |
| 10.3.2 | Maximum likelihood estimation | 360 |
| 10.3.3 | Gradient-based optimization | 363 |

| | | |
|-----------|--|------------|
| 10.3.4 | Bound optimization | 363 |
| 10.3.5 | MAP estimation | 364 |
| 10.3.6 | Maximum entropy classifiers | 365 |
| 10.3.7 | Hierarchical classification | 366 |
| 10.3.8 | Handling large numbers of classes | 367 |
| 10.4 | Robust logistic regression * | 368 |
| 10.4.1 | Mixture model for the likelihood | 368 |
| 10.4.2 | Bi-tempered loss | 369 |
| 10.5 | Bayesian logistic regression * | 371 |
| 10.5.1 | Laplace approximation | 372 |
| 10.5.2 | Approximating the posterior predictive | 374 |
| 10.6 | Exercises | 376 |
| 11 | Linear regression | 379 |
| 11.1 | Introduction | 379 |
| 11.2 | Standard linear regression | 379 |
| 11.2.1 | Terminology | 379 |
| 11.2.2 | Least squares estimation | 380 |
| 11.2.3 | Other approaches to computing the MLE | 384 |
| 11.2.4 | Measuring goodness of fit | 387 |
| 11.3 | Ridge regression | 389 |
| 11.3.1 | Computing the MAP estimate | 389 |
| 11.3.2 | Connection between ridge regression and PCA | 391 |
| 11.3.3 | Choosing the strength of the regularizer | 393 |
| 11.4 | Robust linear regression * | 393 |
| 11.4.1 | Robust regression using the Student t distribution | 394 |
| 11.4.2 | Robust regression using the Laplace distribution | 395 |
| 11.4.3 | Robust regression using Huber loss | 396 |
| 11.4.4 | Robust regression by randomly or iteratively removing outliers | 397 |
| 11.5 | Lasso regression | 397 |
| 11.5.1 | MAP estimation with a Laplace prior (ℓ_1 regularization) | 397 |
| 11.5.2 | Why does ℓ_1 regularization yield sparse solutions? | 398 |
| 11.5.3 | Hard vs soft thresholding | 399 |
| 11.5.4 | Regularization path | 401 |
| 11.5.5 | Comparison of least squares, lasso, ridge and subset selection | 402 |
| 11.5.6 | Variable selection consistency | 404 |
| 11.5.7 | Group lasso | 405 |
| 11.5.8 | Elastic net (ridge and lasso combined) | 407 |
| 11.5.9 | Optimization algorithms | 407 |
| 11.6 | Bayesian linear regression * | 409 |
| 11.6.1 | Computing the posterior | 409 |
| 11.6.2 | Computing the posterior predictive | 412 |
| 11.6.3 | Empirical Bayes (Automatic relevancy determination) | 414 |
| 11.7 | Exercises | 416 |

| | |
|--|------------|
| 12 Generalized linear models * | 419 |
| 12.1 Introduction | 419 |
| 12.2 The exponential family | 419 |
| 12.2.1 Definition | 419 |
| 12.2.2 Examples | 420 |
| 12.2.3 Log partition function is cumulant generating function | 425 |
| 12.2.4 MLE for the exponential family | 426 |
| 12.2.5 Exponential dispersion family | 427 |
| 12.2.6 Maximum entropy derivation of the exponential family | 427 |
| 12.3 Generalized linear models (GLMs) | 428 |
| 12.3.1 Examples | 429 |
| 12.3.2 Maximum likelihood estimation | 431 |
| 12.3.3 GLMs with non-canonical link functions | 431 |
| 12.4 Probit regression | 432 |
| 12.4.1 Latent variable interpretation | 432 |
| 12.4.2 Maximum likelihood estimation | 433 |
| 12.4.3 Ordinal probit regression * | 435 |
| 12.4.4 Multinomial probit models * | 435 |
| III Deep neural networks | 437 |
| 13 Neural networks for unstructured data | 439 |
| 13.1 Introduction | 439 |
| 13.2 Multilayer perceptrons (MLPs) | 440 |
| 13.2.1 The XOR problem | 440 |
| 13.2.2 Differentiable MLPs | 442 |
| 13.2.3 Activation functions | 442 |
| 13.2.4 Example models | 443 |
| 13.2.5 The importance of depth | 448 |
| 13.2.6 Connections with biology | 449 |
| 13.3 Backpropagation | 451 |
| 13.3.1 Forward vs reverse mode differentiation | 452 |
| 13.3.2 Reverse mode differentiation for multilayer perceptrons | 453 |
| 13.3.3 Vector-Jacobian product for common layers | 455 |
| 13.3.4 Computation graphs | 457 |
| 13.3.5 Automatic differentiation in functional form * | 460 |
| 13.4 Training neural networks | 464 |
| 13.4.1 Tuning the learning rate | 464 |
| 13.4.2 Vanishing and exploding gradients | 464 |
| 13.4.3 Non-saturating activation functions | 465 |
| 13.4.4 Residual connections | 468 |
| 13.4.5 Parameter initialization | 469 |
| 13.4.6 Multi-GPU training | 472 |
| 13.5 Regularization | 473 |

| | | | |
|-----------|---|------------|--|
| 13.5.1 | Early stopping | 473 | |
| 13.5.2 | Weight decay | 473 | |
| 13.5.3 | Sparse DNNs | 473 | |
| 13.5.4 | Dropout | 474 | |
| 13.5.5 | Bayesian neural networks | 475 | |
| 13.5.6 | Regularization effects of (stochastic) gradient descent * | 475 | |
| 13.6 | Other kinds of feedforward networks | 477 | |
| 13.6.1 | Radial basis function networks | 477 | |
| 13.6.2 | Mixtures of experts | 479 | |
| 13.7 | Exercises | 482 | |
| 14 | Neural networks for images | 485 | |
| 14.1 | Introduction | 485 | |
| 14.2 | Common layers | 486 | |
| 14.2.1 | Convolutional layers | 486 | |
| 14.2.2 | Pooling layers | 493 | |
| 14.2.3 | Putting it altogether | 493 | |
| 14.2.4 | Normalization layers | 494 | |
| 14.3 | Common architectures for image classification | 497 | |
| 14.3.1 | LeNet | 497 | |
| 14.3.2 | AlexNet | 498 | |
| 14.3.3 | GoogLeNet (Inception) | 499 | |
| 14.3.4 | ResNet | 500 | |
| 14.3.5 | DenseNet | 501 | |
| 14.3.6 | Neural architecture search | 502 | |
| 14.4 | Other forms of convolution * | 502 | |
| 14.4.1 | Dilated convolution | 502 | |
| 14.4.2 | Transposed convolution | 503 | |
| 14.4.3 | Depthwise separable convolution | 504 | |
| 14.5 | Solving other discriminative vision tasks with CNNs | 505 | |
| 14.5.1 | Image tagging | 505 | |
| 14.5.2 | Object detection | 506 | |
| 14.5.3 | Instance segmentation | 507 | |
| 14.5.4 | Semantic segmentation | 507 | |
| 14.5.5 | Human pose estimation | 508 | |
| 14.6 | Generating images by inverting CNNs * | 509 | |
| 14.6.1 | Converting a trained classifier into a generative model | 509 | |
| 14.6.2 | Image priors | 510 | |
| 14.6.3 | Visualizing the features learned by a CNN | 511 | |
| 14.6.4 | Deep Dream | 512 | |
| 14.6.5 | Neural style transfer | 513 | |
| 14.7 | Adversarial Examples * | 516 | |
| 14.7.1 | Whitebox (gradient-based) attacks | 517 | |
| 14.7.2 | Blackbox (gradient-free) attacks | 518 | |
| 14.7.3 | Real world adversarial attacks | 519 | |

| | | |
|-----------|--|------------|
| 14.7.4 | Defenses based on robust optimization | 520 |
| 14.7.5 | Why models have adversarial examples | 521 |
| 15 | Neural networks for sequences | 523 |
| 15.1 | Introduction | 523 |
| 15.2 | Recurrent neural networks (RNNs) | 523 |
| 15.2.1 | Vec2Seq (sequence generation) | 523 |
| 15.2.2 | Seq2Vec (sequence classification) | 526 |
| 15.2.3 | Seq2Seq (sequence translation) | 527 |
| 15.2.4 | Teacher forcing | 529 |
| 15.2.5 | Backpropagation through time | 530 |
| 15.2.6 | Vanishing and exploding gradients | 531 |
| 15.2.7 | Gating and long term memory | 532 |
| 15.2.8 | Beam search | 535 |
| 15.3 | 1d CNNs | 536 |
| 15.3.1 | 1d CNNs for sequence classification | 536 |
| 15.3.2 | Causal 1d CNNs for sequence generation | 537 |
| 15.4 | Attention | 538 |
| 15.4.1 | Attention as soft dictionary lookup | 538 |
| 15.4.2 | Kernel regression as non-parametric attention | 539 |
| 15.4.3 | Parametric attention | 540 |
| 15.4.4 | Seq2Seq with attention | 541 |
| 15.4.5 | Seq2vec with attention (text classification) | 543 |
| 15.4.6 | Seq+Seq2Vec with attention (text pair classification) | 543 |
| 15.4.7 | Soft vs hard attention | 545 |
| 15.5 | Transformers | 546 |
| 15.5.1 | Self-attention | 546 |
| 15.5.2 | Multi-headed attention | 546 |
| 15.5.3 | Positional encoding | 547 |
| 15.5.4 | Putting it altogether | 549 |
| 15.5.5 | Comparing transformers, CNNs and RNNs | 550 |
| 15.5.6 | Transformers for images * | 551 |
| 15.6 | Efficient transformers * | 552 |
| 15.6.1 | Fixed non-learnable localized attention patterns | 552 |
| 15.6.2 | Learnable sparse attention patterns | 553 |
| 15.6.3 | Memory and recurrence methods | 554 |
| 15.6.4 | Low-rank and kernel methods | 554 |
| 15.7 | Language models and unsupervised representation learning | 556 |
| 15.7.1 | ELMo | 557 |
| 15.7.2 | BERT | 557 |
| 15.7.3 | GPT | 561 |
| 15.7.4 | T5 | 562 |
| 15.7.5 | Discussion | 562 |

IV Nonparametric models 563

16 Exemplar-based methods 565

- 16.1 K nearest neighbor (KNN) classification 565
 - 16.1.1 Example 566
 - 16.1.2 The curse of dimensionality 566
 - 16.1.3 Reducing the speed and memory requirements 568
 - 16.1.4 Open set recognition 568
- 16.2 Learning distance metrics 569
 - 16.2.1 Linear and convex methods 570
 - 16.2.2 Deep metric learning 571
 - 16.2.3 Classification losses 572
 - 16.2.4 Ranking losses 572
 - 16.2.5 Speeding up ranking loss optimization 574
 - 16.2.6 Other training tricks for DML 577
- 16.3 Kernel density estimation (KDE) 577
 - 16.3.1 Density kernels 578
 - 16.3.2 Parzen window density estimator 578
 - 16.3.3 How to choose the bandwidth parameter 580
 - 16.3.4 From KDE to KNN classification 580
 - 16.3.5 Kernel regression 581

17 Kernel methods 585

- 17.1 Inferring functions from data 585
 - 17.1.1 Smoothness prior 586
 - 17.1.2 Inference from noise-free observations 586
 - 17.1.3 Inference from noisy observations 588
- 17.2 Mercer kernels 588
 - 17.2.1 Mercer's theorem 589
 - 17.2.2 Some popular Mercer kernels 589
- 17.3 Gaussian processes 594
 - 17.3.1 Noise-free observations 594
 - 17.3.2 Noisy observations 595
 - 17.3.3 Comparison to kernel regression 596
 - 17.3.4 Weight space vs function space 597
 - 17.3.5 Numerical issues 598
 - 17.3.6 Estimating the kernel 598
 - 17.3.7 GPs for classification 601
 - 17.3.8 Connections with deep learning 602
- 17.4 Scaling GPs to large datasets 603
 - 17.4.1 Sparse (inducing-point) approximations 603
 - 17.4.2 Exploiting parallelization and kernel matrix structure 603
 - 17.4.3 Random feature approximation 603
- 17.5 Support vector machines (SVMs) 605
 - 17.5.1 Large margin classifiers 605

| | | | |
|-----------|---|------------|--|
| 17.5.2 | The dual problem | 607 | |
| 17.5.3 | Soft margin classifiers | 609 | |
| 17.5.4 | The kernel trick | 610 | |
| 17.5.5 | Converting SVM outputs into probabilities | 611 | |
| 17.5.6 | Connection with logistic regression | 611 | |
| 17.5.7 | Multi-class classification with SVMs | 612 | |
| 17.5.8 | How to choose the regularizer C | 613 | |
| 17.5.9 | Kernel ridge regression | 614 | |
| 17.5.10 | SVMs for regression | 615 | |
| 17.6 | Sparse vector machines | 617 | |
| 17.6.1 | Relevance vector machines (RVMs) | 618 | |
| 17.6.2 | Comparison of sparse and dense kernel methods | 618 | |
| 17.7 | Optimizing in function space * | 620 | |
| 17.7.1 | Functional analysis | 621 | |
| 17.7.2 | Hilbert space | 622 | |
| 17.7.3 | Reproducing Kernel Hilbert Space | 622 | |
| 17.7.4 | Representer theorem | 623 | |
| 17.7.5 | Kernel ridge regression revisited | 625 | |
| 17.8 | Exercises | 625 | |
| 18 | Trees, forests, bagging and boosting | 627 | |
| 18.1 | Classification and regression trees (CART) | 627 | |
| 18.1.1 | Model definition | 627 | |
| 18.1.2 | Model fitting | 629 | |
| 18.1.3 | Regularization | 630 | |
| 18.1.4 | Handling missing input features | 630 | |
| 18.1.5 | Pros and cons | 630 | |
| 18.2 | Ensemble learning | 632 | |
| 18.2.1 | Stacking | 632 | |
| 18.2.2 | Ensembling is not Bayes model averaging | 633 | |
| 18.3 | Bagging | 633 | |
| 18.4 | Random forests | 634 | |
| 18.5 | Boosting | 635 | |
| 18.5.1 | Forward stagewise additive modeling | 636 | |
| 18.5.2 | Quadratic loss and least squares boosting | 636 | |
| 18.5.3 | Exponential loss and AdaBoost | 637 | |
| 18.5.4 | LogitBoost | 640 | |
| 18.5.5 | Gradient boosting | 640 | |
| 18.6 | Interpreting tree ensembles | 644 | |
| 18.6.1 | Feature importance | 644 | |
| 18.6.2 | Partial dependency plots | 645 | |

| | | |
|-----------|--|------------|
| V | Beyond supervised learning | 647 |
| 19 | Learning with fewer labeled examples | 649 |
| 19.1 | Data augmentation | 649 |
| 19.1.1 | Examples | 649 |
| 19.1.2 | Theoretical justification | 650 |
| 19.2 | Transfer learning | 650 |
| 19.2.1 | Fine-tuning | 651 |
| 19.2.2 | Supervised pre-training | 652 |
| 19.2.3 | Unsupervised pre-training (self-supervised learning) | 653 |
| 19.2.4 | Domain adaptation | 656 |
| 19.3 | Semi-supervised learning | 656 |
| 19.3.1 | Self-training and pseudo-labeling | 657 |
| 19.3.2 | Entropy minimization | 658 |
| 19.3.3 | Co-training | 661 |
| 19.3.4 | Label propagation on graphs | 661 |
| 19.3.5 | Consistency regularization | 662 |
| 19.3.6 | Deep generative models * | 664 |
| 19.3.7 | Combining self-supervised and semi-supervised learning | 668 |
| 19.4 | Active learning | 668 |
| 19.4.1 | Decision-theoretic approach | 669 |
| 19.4.2 | Information-theoretic approach | 669 |
| 19.4.3 | Batch active learning | 670 |
| 19.5 | Meta-learning * | 670 |
| 19.5.1 | Model-agnostic meta-learning (MAML) | 670 |
| 19.6 | Few-shot learning * | 671 |
| 19.6.1 | Matching networks | 672 |
| 19.7 | Exercises | 673 |
| 20 | Dimensionality reduction | 675 |
| 20.1 | Principal components analysis (PCA) | 675 |
| 20.1.1 | Examples | 675 |
| 20.1.2 | Derivation of the algorithm | 677 |
| 20.1.3 | Computational issues | 680 |
| 20.1.4 | Choosing the number of latent dimensions | 682 |
| 20.2 | Factor analysis * | 684 |
| 20.2.1 | Generative model | 685 |
| 20.2.2 | Probabilistic PCA | 686 |
| 20.2.3 | EM algorithm for FA/PPCA | 687 |
| 20.2.4 | Unidentifiability of the parameters | 689 |
| 20.2.5 | Nonlinear factor analysis | 691 |
| 20.2.6 | Mixtures of factor analysers | 692 |
| 20.2.7 | Exponential family factor analysis | 693 |
| 20.2.8 | Factor analysis models for paired data | 695 |
| 20.3 | Autoencoders | 698 |

| | | | |
|-----------|---|------------|--|
| 20.3.1 | Bottleneck autoencoders | 698 | |
| 20.3.2 | Denoising autoencoders | 699 | |
| 20.3.3 | Contractive autoencoders | 700 | |
| 20.3.4 | Sparse autoencoders | 701 | |
| 20.3.5 | Variational autoencoders | 703 | |
| 20.4 | Manifold learning * | 706 | |
| 20.4.1 | What are manifolds? | 707 | |
| 20.4.2 | The manifold hypothesis | 708 | |
| 20.4.3 | Approaches to manifold learning | 708 | |
| 20.4.4 | Multi-dimensional scaling (MDS) | 709 | |
| 20.4.5 | Isomap | 712 | |
| 20.4.6 | Kernel PCA | 713 | |
| 20.4.7 | Maximum variance unfolding (MVU) | 715 | |
| 20.4.8 | Local linear embedding (LLE) | 715 | |
| 20.4.9 | Laplacian eigenmaps | 717 | |
| 20.4.10 | t-SNE | 719 | |
| 20.5 | Word embeddings | 723 | |
| 20.5.1 | Latent semantic analysis / indexing | 723 | |
| 20.5.2 | Word2vec | 725 | |
| 20.5.3 | GloVE | 728 | |
| 20.5.4 | Word analogies | 728 | |
| 20.5.5 | RAND-WALK model of word embeddings | 729 | |
| 20.5.6 | Contextual word embeddings | 730 | |
| 20.6 | Exercises | 730 | |
| 21 | Clustering | 733 | |
| 21.1 | Introduction | 733 | |
| 21.1.1 | Evaluating the output of clustering methods | 733 | |
| 21.2 | Hierarchical agglomerative clustering | 735 | |
| 21.2.1 | The algorithm | 736 | |
| 21.2.2 | Example | 738 | |
| 21.3 | K means clustering | 739 | |
| 21.3.1 | The algorithm | 740 | |
| 21.3.2 | Examples | 740 | |
| 21.3.3 | Vector quantization | 741 | |
| 21.3.4 | The K-means++ algorithm | 743 | |
| 21.3.5 | The K-medoids algorithm | 743 | |
| 21.3.6 | Speedup tricks | 744 | |
| 21.3.7 | Choosing the number of clusters K | 744 | |
| 21.4 | Clustering using mixture models | 748 | |
| 21.4.1 | Mixtures of Gaussians | 748 | |
| 21.4.2 | Mixtures of Bernoullis | 752 | |
| 21.5 | Spectral clustering * | 752 | |
| 21.5.1 | Normalized cuts | 752 | |
| 21.5.2 | Eigenvectors of the graph Laplacian encode the clustering | 753 | |

| | | | |
|-------------------|---|------------|--|
| 21.5.3 | Example | 754 | |
| 21.5.4 | Connection with other methods | 754 | |
| 21.6 | Biclustering * | 755 | |
| 21.6.1 | Basic biclustering | 755 | |
| 21.6.2 | Nested partition models (Crosscat) | 756 | |
| 22 | Recommender systems | 759 | |
| 22.1 | Explicit feedback | 759 | |
| 22.1.1 | Datasets | 759 | |
| 22.1.2 | Collaborative filtering | 760 | |
| 22.1.3 | Matrix factorization | 761 | |
| 22.1.4 | Autoencoders | 763 | |
| 22.2 | Implicit feedback | 764 | |
| 22.2.1 | Bayesian personalized ranking | 765 | |
| 22.2.2 | Factorization machines | 765 | |
| 22.2.3 | Neural matrix factorization | 766 | |
| 22.3 | Leveraging side information | 767 | |
| 22.4 | Exploration-exploitation tradeoff | 768 | |
| 23 | Graph embeddings * | 771 | |
| 23.1 | Introduction | 771 | |
| 23.2 | Graph Embedding as an Encoder/Decoder Problem | 772 | |
| 23.3 | Shallow graph embeddings | 774 | |
| 23.3.1 | Unsupervised embeddings | 775 | |
| 23.3.2 | Distance-based: Euclidean methods | 775 | |
| 23.3.3 | Distance-based: non-Euclidean methods | 776 | |
| 23.3.4 | Outer product-based: Matrix factorization methods | 776 | |
| 23.3.5 | Outer product-based: Skip-gram methods | 777 | |
| 23.3.6 | Supervised embeddings | 778 | |
| 23.4 | Graph Neural Networks | 779 | |
| 23.4.1 | Message passing GNNs | 779 | |
| 23.4.2 | Spectral Graph Convolutions | 781 | |
| 23.4.3 | Spatial Graph Convolutions | 781 | |
| 23.4.4 | Non-Euclidean Graph Convolutions | 783 | |
| 23.5 | Deep graph embeddings | 783 | |
| 23.5.1 | Unsupervised embeddings | 783 | |
| 23.5.2 | Semi-supervised embeddings | 786 | |
| 23.6 | Applications | 787 | |
| 23.6.1 | Unsupervised applications | 787 | |
| 23.6.2 | Supervised applications | 789 | |
| Appendices | | 791 | |

VI Appendix 793

A Notation 795

| | | | |
|-------|---|-----|--|
| A.1 | Introduction | 795 | |
| A.2 | Common mathematical symbols | 795 | |
| A.3 | Functions | 796 | |
| A.3.1 | Common functions of one argument | 796 | |
| A.3.2 | Common functions of two arguments | 796 | |
| A.3.3 | Common functions of > 2 arguments | 796 | |
| A.4 | Linear algebra | 797 | |
| A.4.1 | General notation | 797 | |
| A.4.2 | Vectors | 797 | |
| A.4.3 | Matrices | 797 | |
| A.4.4 | Matrix calculus | 798 | |
| A.5 | Optimization | 798 | |
| A.6 | Probability | 799 | |
| A.7 | Information theory | 799 | |
| A.8 | Statistics and machine learning | 799 | |
| A.8.1 | Supervised learning | 800 | |
| A.8.2 | Unsupervised learning and generative models | 800 | |
| A.8.3 | Bayesian inference | 800 | |
| A.9 | Abbreviations | 801 | |

Bibliography 815