

Contents

Preface	xxvii
----------------	--------------

List of Figures	xxix
------------------------	-------------

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	45
2.4.2	Sigmoid (logistic) function	47
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 *	61
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	67
2.8.8	Probability integral transform	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	The exponential family	84
3.4.1	Definition	84
3.4.2	Example	85
3.4.3	Log partition function is cumulant generating function	86
3.4.4	Maximum entropy derivation of the exponential family	86
3.5	Mixture models	87
3.5.1	Gaussian mixture models	88
3.5.2	Mixtures of Bernoullis	89
3.5.3	Gaussian scale mixtures *	90
3.6	Probabilistic graphical models *	91
3.6.1	Representation	91
3.6.2	Inference	94
3.6.3	Learning	94
3.7	Exercises	96
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	122
4.6.2	The beta-binomial model	122
4.6.3	The Dirichlet-multinomial model	129
4.6.4	The Gaussian-Gaussian model	133
4.6.5	Beyond conjugate priors	136
4.6.6	Credible intervals	138
4.6.7	Bayesian machine learning	139
4.6.8	Computational issues	143
4.7	Frequentist statistics *	146
4.7.1	Sampling distributions	146
4.7.2	Gaussian approximation of the sampling distribution of the MLE	147
4.7.3	Bootstrap approximation of the sampling distribution of any estimator	148
4.7.4	Confidence intervals	149
4.7.5	Caution: Confidence intervals are not credible	150
4.7.6	The bias-variance tradeoff	151
4.8	Exercises	156
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
5.1.7	Information criteria	171
5.2	Bayesian hypothesis testing	173
5.2.1	Example: Testing if a coin is fair	173
5.2.2	Bayesian model selection	174
5.2.3	Occam's razor	176
5.2.4	Connection between cross validation and marginal likelihood	177
5.3	Frequentist decision theory	178
5.3.1	Computing the risk of an estimator	178
5.3.2	Consistent estimators	181
5.3.3	Admissible estimators	181

5.4	Empirical risk minimization	182
5.4.1	Empirical risk	182
5.4.2	Structural risk	184
5.4.3	Cross-validation	185
5.4.4	Statistical learning theory *	186
5.5	Frequentist hypothesis testing *	187
5.5.1	Likelihood ratio test	187
5.5.2	Null hypothesis significance testing (NHST)	188
5.5.3	p-values	189
5.5.4	p-values considered harmful	189
5.5.5	Why isn't everyone a Bayesian?	192
5.6	Exercises	193
6	Information theory	195
6.1	Entropy	195
6.1.1	Entropy for discrete random variables	195
6.1.2	Cross entropy	197
6.1.3	Joint entropy	197
6.1.4	Conditional entropy	198
6.1.5	Perplexity	199
6.1.6	Differential entropy for continuous random variables *	200
6.2	Relative entropy (KL divergence) *	201
6.2.1	Definition	201
6.2.2	Interpretation	202
6.2.3	Example: KL divergence between two Gaussians	202
6.2.4	Non-negativity of KL	202
6.2.5	KL divergence and MLE	203
6.2.6	Forward vs reverse KL	204
6.3	Mutual information *	205
6.3.1	Definition	205
6.3.2	Interpretation	206
6.3.3	Example	206
6.3.4	Conditional mutual information	207
6.3.5	MI as a “generalized correlation coefficient”	208
6.3.6	Normalized mutual information	209
6.3.7	Maximal information coefficient	209
6.3.8	Data processing inequality	211
6.3.9	Sufficient Statistics	212
6.3.10	Fano's inequality *	213
6.4	Exercises	214
7	Linear algebra	217
7.1	Introduction	217
7.1.1	Notation	217
7.1.2	Vector spaces	220

7.1.3	Norms of a vector and matrix	222	
7.1.4	Properties of a matrix	224	
7.1.5	Special types of matrices	226	
7.2	Matrix multiplication	230	
7.2.1	Vector-Vector Products	230	
7.2.2	Matrix-Vector Products	230	
7.2.3	Matrix-Matrix Products	231	
7.2.4	Application: manipulating data matrices	233	
7.2.5	Kronecker products *	235	
7.2.6	Einstein summation *	236	
7.3	Matrix inversion	237	
7.3.1	The inverse of a square matrix	237	
7.3.2	Schur complements *	237	
7.3.3	The matrix inversion lemma *	239	
7.3.4	Matrix determinant lemma *	239	
7.4	Eigenvalue decomposition (EVD)	240	
7.4.1	Basics	240	
7.4.2	Diagonalization	241	
7.4.3	Eigenvalues and eigenvectors of symmetric matrices	241	
7.4.4	Geometry of quadratic forms	242	
7.4.5	Standardizing and whitening data	242	
7.4.6	Power method	244	
7.4.7	Deflation	245	
7.4.8	Eigenvectors optimize quadratic forms	245	
7.5	Singular value decomposition (SVD)	245	
7.5.1	Basics	245	
7.5.2	Connection between SVD and EVD	246	
7.5.3	Pseudo inverse	247	
7.5.4	SVD and the range and null space of a matrix *	248	
7.5.5	Truncated SVD	249	
7.6	Other matrix decompositions *	250	
7.6.1	LU factorization	250	
7.6.2	QR decomposition	250	
7.6.3	Cholesky decomposition	251	
7.7	Solving systems of linear equations *	252	
7.7.1	Solving square systems	253	
7.7.2	Solving underconstrained systems (least norm estimation)	253	
7.7.3	Solving overconstrained systems (least squares estimation)	254	
7.8	Matrix calculus	255	
7.8.1	Derivatives	255	
7.8.2	Gradients	256	
7.8.3	Directional derivative	257	
7.8.4	Total derivative *	257	
7.8.5	Jacobian	257	
7.8.6	Hessian	258	

7.8.7	Gradients of commonly used functions	258
7.8.8	Functional derivative notation *	260
7.9	Exercises	263
8	Optimization	265
8.1	Introduction	265
8.1.1	Local vs global optimization	265
8.1.2	Constrained vs unconstrained optimization	267
8.1.3	Convex vs nonconvex optimization	267
8.1.4	Smooth vs nonsmooth optimization	271
8.2	First-order methods	272
8.2.1	Descent direction	273
8.2.2	Step size (learning rate)	274
8.2.3	Convergence rates	276
8.2.4	Momentum methods	277
8.3	Second-order methods	278
8.3.1	Newton's method	279
8.3.2	BFGS and other quasi-Newton methods	280
8.3.3	Trust region methods	281
8.3.4	Natural gradient descent *	282
8.4	Stochastic gradient descent	285
8.4.1	Application to finite sum problems	285
8.4.2	Example: SGD for fitting linear regression	286
8.4.3	Choosing the step size (learning rate)	287
8.4.4	Iterate averaging	289
8.4.5	Variance reduction *	289
8.4.6	Preconditioned SGD	291
8.5	Constrained optimization	293
8.5.1	Lagrange multipliers	294
8.5.2	The KKT conditions	295
8.5.3	Linear programming	297
8.5.4	Quadratic programming	298
8.5.5	Mixed integer linear programming *	299
8.6	Proximal gradient method *	300
8.6.1	Projected gradient descent	300
8.6.2	Proximal operator for ℓ_1 -norm regularizer	301
8.6.3	Proximal operator for quantization	303
8.7	Bound optimization *	303
8.7.1	The general algorithm	304
8.7.2	The EM algorithm	304
8.7.3	Example: EM for a GMM	307
8.7.4	Example: EM for an MVN with missing data	311
8.8	Blackbox and derivative free optimization	314
8.8.1	Grid search and random search	314
8.8.2	Simulated annealing *	314

8.8.3	Model-based blackbox optimization *	315
8.9	Exercises	316
II	Linear models	317
9	Linear discriminant analysis	319
9.1	Introduction	319
9.2	Gaussian discriminant analysis	319
9.2.1	Quadratic decision boundaries	320
9.2.2	Linear decision boundaries	321
9.2.3	The connection between LDA and logistic regression	321
9.2.4	Model fitting	322
9.2.5	Nearest centroid classifier	324
9.2.6	Fisher's linear discriminant analysis *	324
9.3	Naive Bayes classifiers	328
9.3.1	Example models	328
9.3.2	Model fitting	329
9.3.3	Bayesian naive Bayes	330
9.3.4	The connection between naive Bayes and logistic regression	331
9.4	Generative vs discriminative classifiers	332
9.4.1	Advantages of discriminative classifiers	332
9.4.2	Advantages of generative classifiers	333
9.4.3	Handling missing features	333
9.5	Exercises	334
10	Logistic regression	335
10.1	Introduction	335
10.2	Binary logistic regression	335
10.2.1	Linear classifiers	335
10.2.2	Nonlinear classifiers	336
10.2.3	Maximum likelihood estimation	338
10.2.4	Stochastic gradient descent	341
10.2.5	Perceptron algorithm	341
10.2.6	Iteratively reweighted least squares	342
10.2.7	MAP estimation	343
10.2.8	Standardization	345
10.3	Multinomial logistic regression	346
10.3.1	Linear and nonlinear classifiers	346
10.3.2	Maximum likelihood estimation	346
10.3.3	Gradient-based optimization	349
10.3.4	Bound optimization	349
10.3.5	MAP estimation	351
10.3.6	Maximum entropy classifiers	351
10.3.7	Hierarchical classification	352

10.3.8	Handling large numbers of classes	353
10.4	Robust logistic regression *	355
10.4.1	Mixture model for the likelihood	355
10.4.2	Bi-tempered loss	356
10.5	Bayesian logistic regression *	358
10.5.1	Laplace approximation	358
10.5.2	Approximating the posterior predictive	360
10.6	Exercises	362
11	Linear regression	365
11.1	Introduction	365
11.2	Least squares linear regression	365
11.2.1	Terminology	365
11.2.2	Least squares estimation	366
11.2.3	Other approaches to computing the MLE	370
11.2.4	Measuring goodness of fit	374
11.3	Ridge regression	375
11.3.1	Computing the MAP estimate	376
11.3.2	Connection between ridge regression and PCA	378
11.3.3	Choosing the strength of the regularizer	379
11.4	Robust linear regression *	379
11.4.1	Student- t likelihood	380
11.4.2	Laplace likelihood	381
11.4.3	Huber loss	382
11.4.4	RANSAC	383
11.5	Lasso regression	383
11.5.1	MAP estimation with a Laplace prior (ℓ_1 regularization)	384
11.5.2	Why does ℓ_1 regularization yield sparse solutions?	384
11.5.3	Hard vs soft thresholding	385
11.5.4	Regularization path	387
11.5.5	Comparison of least squares, lasso, ridge and subset selection	389
11.5.6	Variable selection consistency	389
11.5.7	Group lasso	391
11.5.8	Elastic net (ridge and lasso combined)	393
11.5.9	Optimization algorithms	394
11.6	Bayesian linear regression *	396
11.6.1	Computing the posterior	396
11.6.2	Computing the posterior predictive	399
11.6.3	Empirical Bayes (Automatic relevancy determination)	399
11.7	Exercises	402
12	Generalized linear models *	405
12.1	Introduction	405
12.2	Examples	405
12.2.1	Linear regression	406

12.2.2	Binomial regression	406	
12.2.3	Poisson regression	407	
12.3	GLMs with non-canonical link functions	407	
12.4	Maximum likelihood estimation	408	
III	Deep neural networks	411	
13	Neural networks for unstructured data	413	
13.1	Introduction	413	
13.2	Multilayer perceptrons (MLPs)	414	
13.2.1	The XOR problem	414	
13.2.2	Differentiable MLPs	415	
13.2.3	Activation functions	416	
13.2.4	Example models	417	
13.2.5	The importance of depth	421	
13.2.6	Connections with biology	423	
13.3	Backpropagation	425	
13.3.1	Forward vs reverse mode differentiation	426	
13.3.2	Reverse mode differentiation for multilayer perceptrons	427	
13.3.3	Vector-Jacobian product for common layers	428	
13.3.4	Computation graphs	431	
13.3.5	Automatic differentiation in functional form *	433	
13.4	Training neural networks	438	
13.4.1	Tuning the learning rate	438	
13.4.2	Vanishing and exploding gradients	438	
13.4.3	Non-saturating activation functions	439	
13.4.4	Residual connections	441	
13.4.5	Parameter initialization	442	
13.4.6	Multi-GPU training	445	
13.5	Regularization	446	
13.5.1	Early stopping	446	
13.5.2	Weight decay	447	
13.5.3	Sparse DNNs	447	
13.5.4	Dropout	447	
13.5.5	Bayesian neural networks	449	
13.5.6	Regularization effects of (stochastic) gradient descent *	449	
13.6	Other kinds of feedforward networks	451	
13.6.1	Radial basis function networks	451	
13.6.2	Mixtures of experts	452	
13.7	Exercises	456	
14	Neural networks for images	457	
14.1	Introduction	457	
14.2	Common layers	458	

14.2.1	Convolutional layers	458	
14.2.2	Pooling layers	465	
14.2.3	Putting it altogether	466	
14.2.4	Normalization layers	466	
14.3	Common architectures for image classification	469	
14.3.1	LeNet	469	
14.3.2	AlexNet	470	
14.3.3	GoogLeNet (Inception)	471	
14.3.4	ResNet	472	
14.3.5	DenseNet	473	
14.3.6	Neural architecture search	474	
14.4	Other forms of convolution *	474	
14.4.1	Dilated convolution	475	
14.4.2	Transposed convolution	475	
14.4.3	Depthwise separable convolution	477	
14.5	Solving other discriminative vision tasks with CNNs	477	
14.5.1	Image tagging	478	
14.5.2	Object detection	478	
14.5.3	Instance segmentation	479	
14.5.4	Semantic segmentation	479	
14.5.5	Human pose estimation	481	
14.6	Generating images by inverting CNNs *	481	
14.6.1	Converting a trained classifier into a generative model	482	
14.6.2	Image priors	482	
14.6.3	Visualizing the features learned by a CNN	483	
14.6.4	Deep Dream	485	
14.6.5	Neural style transfer	485	
14.7	Adversarial Examples *	488	
14.7.1	Whitebox (gradient-based) attacks	489	
14.7.2	Blackbox (gradient-free) attacks	490	
14.7.3	Real world adversarial attacks	491	
14.7.4	Defenses based on robust optimization	492	
14.7.5	Why models have adversarial examples	493	
15	Neural networks for sequences	497	
15.1	Introduction	497	
15.2	Recurrent neural networks (RNNs)	497	
15.2.1	Vec2Seq (sequence generation)	497	
15.2.2	Seq2Vec (sequence classification)	499	
15.2.3	Seq2Seq (sequence translation)	501	
15.2.4	Teacher forcing	503	
15.2.5	Backpropagation through time	504	
15.2.6	Vanishing and exploding gradients	505	
15.2.7	Gating and long term memory	506	
15.2.8	Beam search	509	

15.3	1d CNNs	510	
15.3.1	1d CNNs for sequence classification	510	
15.3.2	Causal 1d CNNs for sequence generation	511	
15.4	Attention	512	
15.4.1	Attention as soft dictionary lookup	512	
15.4.2	Kernel regression as non-parametric attention	513	
15.4.3	Parametric attention	514	
15.4.4	Seq2Seq with attention	515	
15.4.5	Seq2vec with attention (text classification)	517	
15.4.6	Seq+Seq2Vec with attention (text pair classification)	517	
15.4.7	Soft vs hard attention	519	
15.5	Transformers	520	
15.5.1	Self-attention	520	
15.5.2	Multi-headed attention	520	
15.5.3	Positional encoding	521	
15.5.4	Putting it altogether	523	
15.5.5	Comparing transformers, CNNs and RNNs	524	
15.5.6	Transformers for images *	525	
15.6	Efficient transformers *	526	
15.6.1	Fixed non-learnable localized attention patterns	526	
15.6.2	Learnable sparse attention patterns	527	
15.6.3	Memory and recurrence methods	528	
15.6.4	Low-rank and kernel methods	528	
15.7	Language models and unsupervised representation learning	530	
15.7.1	ELMo	531	
15.7.2	BERT	531	
15.7.3	GPT	535	
15.7.4	T5	536	
15.7.5	Discussion	536	

IV Nonparametric models 537

16 Exemplar-based methods 539

16.1	K nearest neighbor (KNN) classification	539
16.1.1	Example	540
16.1.2	The curse of dimensionality	540
16.1.3	Reducing the speed and memory requirements	542
16.1.4	Open set recognition	542
16.2	Learning distance metrics	543
16.2.1	Linear and convex methods	544
16.2.2	Deep metric learning	545
16.2.3	Classification losses	546
16.2.4	Ranking losses	546
16.2.5	Speeding up ranking loss optimization	548

16.2.6	Other training tricks for DML	551
16.3	Kernel density estimation (KDE)	551
16.3.1	Density kernels	552
16.3.2	Parzen window density estimator	552
16.3.3	How to choose the bandwidth parameter	554
16.3.4	From KDE to KNN classification	554
16.3.5	Kernel regression	555
17	Kernel methods	559
17.1	Inferring functions from data	559
17.1.1	Smoothness prior	560
17.1.2	Inference from noise-free observations	560
17.1.3	Inference from noisy observations	562
17.2	Mercer kernels	562
17.2.1	Mercer's theorem	563
17.2.2	Some popular Mercer kernels	563
17.3	Gaussian processes	568
17.3.1	Noise-free observations	568
17.3.2	Noisy observations	569
17.3.3	Comparison to kernel regression	570
17.3.4	Weight space vs function space	571
17.3.5	Numerical issues	571
17.3.6	Estimating the kernel	572
17.3.7	GPs for classification	575
17.3.8	Connections with deep learning	576
17.4	Scaling GPs to large datasets	576
17.4.1	Sparse (inducing-point) approximations	577
17.4.2	Exploiting parallelization and kernel matrix structure	577
17.4.3	Random feature approximation	577
17.5	Support vector machines (SVMs)	579
17.5.1	Large margin classifiers	579
17.5.2	The dual problem	581
17.5.3	Soft margin classifiers	583
17.5.4	The kernel trick	584
17.5.5	Converting SVM outputs into probabilities	585
17.5.6	Connection with logistic regression	585
17.5.7	Multi-class classification with SVMs	586
17.5.8	How to choose the regularizer C	587
17.5.9	Kernel ridge regression	588
17.5.10	SVMs for regression	589
17.6	Sparse vector machines	592
17.6.1	Relevance vector machines (RVMs)	592
17.6.2	Comparison of sparse and dense kernel methods	592
17.7	Optimizing in function space *	595
17.7.1	Functional analysis	595

17.7.2	Hilbert space	596	
17.7.3	Reproducing Kernel Hilbert Space	597	
17.7.4	Representer theorem	597	
17.7.5	Kernel ridge regression revisited	599	
17.8	Exercises	599	
18	Trees, forests, bagging and boosting	601	
18.1	Classification and regression trees (CART)	601	
18.1.1	Model definition	601	
18.1.2	Model fitting	603	
18.1.3	Regularization	604	
18.1.4	Handling missing input features	604	
18.1.5	Pros and cons	604	
18.2	Ensemble learning	606	
18.2.1	Stacking	606	
18.2.2	Ensembling is not Bayes model averaging	607	
18.3	Bagging	607	
18.4	Random forests	608	
18.5	Boosting	609	
18.5.1	Forward stagewise additive modeling	610	
18.5.2	Quadratic loss and least squares boosting	610	
18.5.3	Exponential loss and AdaBoost	611	
18.5.4	LogitBoost	614	
18.5.5	Gradient boosting	614	
18.6	Interpreting tree ensembles	618	
18.6.1	Feature importance	618	
18.6.2	Partial dependency plots	619	
V	Beyond supervised learning	621	
19	Learning with fewer labeled examples	623	
19.1	Data augmentation	623	
19.1.1	Examples	623	
19.1.2	Theoretical justification	624	
19.2	Transfer learning	624	
19.2.1	Fine-tuning	625	
19.2.2	Adapters	626	
19.2.3	Supervised pre-training	627	
19.2.4	Unsupervised pre-training (self-supervised learning)	628	
19.2.5	Domain adaptation	631	
19.3	Semi-supervised learning	632	
19.3.1	Self-training and pseudo-labeling	632	
19.3.2	Entropy minimization	634	
19.3.3	Co-training	636	

19.3.4	Label propagation on graphs	637
19.3.5	Consistency regularization	638
19.3.6	Deep generative models *	639
19.3.7	Combining self-supervised and semi-supervised learning	643
19.4	Active learning	644
19.4.1	Decision-theoretic approach	644
19.4.2	Information-theoretic approach	645
19.4.3	Batch active learning	645
19.5	Meta-learning *	646
19.5.1	Model-agnostic meta-learning (MAML)	646
19.6	Few-shot learning *	647
19.6.1	Matching networks	648
19.7	Exercises	649
20	Dimensionality reduction	651
20.1	Principal components analysis (PCA)	651
20.1.1	Examples	651
20.1.2	Derivation of the algorithm	653
20.1.3	Computational issues	656
20.1.4	Choosing the number of latent dimensions	658
20.2	Factor analysis *	660
20.2.1	Generative model	661
20.2.2	Probabilistic PCA	662
20.2.3	EM algorithm for FA/PPCA	663
20.2.4	Unidentifiability of the parameters	665
20.2.5	Nonlinear factor analysis	667
20.2.6	Mixtures of factor analysers	668
20.2.7	Exponential family factor analysis	669
20.2.8	Factor analysis models for paired data	671
20.3	Autoencoders	673
20.3.1	Bottleneck autoencoders	674
20.3.2	Denoising autoencoders	675
20.3.3	Contractive autoencoders	675
20.3.4	Sparse autoencoders	677
20.3.5	Variational autoencoders	678
20.4	Manifold learning *	682
20.4.1	What are manifolds?	683
20.4.2	The manifold hypothesis	684
20.4.3	Approaches to manifold learning	684
20.4.4	Multi-dimensional scaling (MDS)	685
20.4.5	Isomap	688
20.4.6	Kernel PCA	688
20.4.7	Maximum variance unfolding (MVU)	690
20.4.8	Local linear embedding (LLE)	691
20.4.9	Laplacian eigenmaps	692

20.4.10	t-SNE	695	
20.5	Word embeddings	699	
20.5.1	Latent semantic analysis / indexing	699	
20.5.2	Word2vec	701	
20.5.3	GloVE	703	
20.5.4	Word analogies	704	
20.5.5	RAND-WALK model of word embeddings	705	
20.5.6	Contextual word embeddings	705	
20.6	Exercises	706	
21	Clustering	709	
21.1	Introduction	709	
21.1.1	Evaluating the output of clustering methods	709	
21.2	Hierarchical agglomerative clustering	711	
21.2.1	The algorithm	712	
21.2.2	Example	714	
21.3	K means clustering	715	
21.3.1	The algorithm	716	
21.3.2	Examples	716	
21.3.3	Vector quantization	717	
21.3.4	The K-means++ algorithm	719	
21.3.5	The K-medoids algorithm	719	
21.3.6	Speedup tricks	720	
21.3.7	Choosing the number of clusters K	720	
21.4	Clustering using mixture models	723	
21.4.1	Mixtures of Gaussians	724	
21.4.2	Mixtures of Bernoullis	728	
21.5	Spectral clustering *	728	
21.5.1	Normalized cuts	728	
21.5.2	Eigenvectors of the graph Laplacian encode the clustering	729	
21.5.3	Example	730	
21.5.4	Connection with other methods	730	
21.6	Biclustering *	731	
21.6.1	Basic biclustering	731	
21.6.2	Nested partition models (Crosscat)	732	
22	Recommender systems	735	
22.1	Explicit feedback	735	
22.1.1	Datasets	735	
22.1.2	Collaborative filtering	736	
22.1.3	Matrix factorization	737	
22.1.4	Autoencoders	739	
22.2	Implicit feedback	740	
22.2.1	Bayesian personalized ranking	741	
22.2.2	Factorization machines	741	

22.2.3	Neural matrix factorization	742
22.3	Leveraging side information	743
22.4	Exploration-exploitation tradeoff	744
23	Graph embeddings *	747
23.1	Introduction	747
23.2	Graph Embedding as an Encoder/Decoder Problem	748
23.3	Shallow graph embeddings	750
23.3.1	Unsupervised embeddings	750
23.3.2	Distance-based: Euclidean methods	751
23.3.3	Distance-based: non-Euclidean methods	752
23.3.4	Outer product-based: Matrix factorization methods	752
23.3.5	Outer product-based: Skip-gram methods	753
23.3.6	Supervised embeddings	754
23.4	Graph Neural Networks	755
23.4.1	Message passing GNNs	755
23.4.2	Spectral Graph Convolutions	757
23.4.3	Spatial Graph Convolutions	757
23.4.4	Non-Euclidean Graph Convolutions	759
23.5	Deep graph embeddings	759
23.5.1	Unsupervised embeddings	759
23.5.2	Semi-supervised embeddings	762
23.6	Applications	763
23.6.1	Unsupervised applications	763
23.6.2	Supervised applications	765
Appendices	767	

VI Appendix 769

A	Notation	771
A.1	Introduction	771
A.2	Common mathematical symbols	771
A.3	Functions	772
A.3.1	Common functions of one argument	772
A.3.2	Common functions of two arguments	772
A.3.3	Common functions of > 2 arguments	772
A.4	Linear algebra	773
A.4.1	General notation	773
A.4.2	Vectors	773
A.4.3	Matrices	773
A.4.4	Matrix calculus	774
A.5	Optimization	774
A.6	Probability	775
A.7	Information theory	775

A.8	Statistics and machine learning	775
A.8.1	Supervised learning	776
A.8.2	Unsupervised learning and generative models	776
A.8.3	Bayesian inference	776
A.9	Abbreviations	777
Bibliography	790	