

Contents

Preface xxvii

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 14
 - 1.3.1 Clustering 14
 - 1.3.2 Discovering latent “factors of variation” 15
 - 1.3.3 Self-supervised learning 16
 - 1.3.4 Evaluating unsupervised learning 16
- 1.4 Reinforcement learning 17
- 1.5 Data 19
 - 1.5.1 Some common image datasets 19
 - 1.5.2 Some common text datasets 21
 - 1.5.3 Preprocessing discrete input data 23
 - 1.5.4 Preprocessing text data 24
 - 1.5.5 Handling missing data 26
- 1.6 Discussion 27
 - 1.6.1 The relationship between ML and other fields 27
 - 1.6.2 Structure of the book 28
 - 1.6.3 Caveats 28

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	
2.2.6	Limitations of summary statistics *	41	
2.3	Bayes' rule	43	
2.3.1	Example: Testing for COVID-19	44	
2.3.2	Example: The Monty Hall problem	45	
2.3.3	Inverse problems *	47	
2.4	Bernoulli and binomial distributions	47	
2.4.1	Definition	47	
2.4.2	Sigmoid (logistic) function	48	
2.4.3	Binary logistic regression	50	
2.5	Categorical and multinomial distributions	51	
2.5.1	Definition	51	
2.5.2	Softmax function	52	
2.5.3	Multiclass logistic regression	53	
2.5.4	Log-sum-exp trick	54	
2.6	Univariate Gaussian (normal) distribution	55	
2.6.1	Cumulative distribution function	55	
2.6.2	Probability density function	56	
2.6.3	Regression	57	
2.6.4	Why is the Gaussian distribution so widely used?	58	
2.6.5	Dirac delta function as a limiting case	58	
2.7	Some other common univariate distributions *	59	
2.7.1	Student t distribution	59	
2.7.2	Cauchy distribution	60	
2.7.3	Laplace distribution	61	
2.7.4	Beta distribution	61	
2.7.5	Gamma distribution	62	
2.7.6	Empirical distribution	63	
2.8	Transformations of random variables *	64	
2.8.1	Discrete case	64	
2.8.2	Continuous case	64	
2.8.3	Invertible transformations (bijections)	65	
2.8.4	Moments of a linear transformation	67	
2.8.5	The convolution theorem	68	
2.8.6	Central limit theorem	69	
2.8.7	Monte Carlo approximation	70	
2.9	Exercises	71	

3	Probability: Multivariate Models	75	
3.1	Joint distributions for multiple random variables	75	
3.1.1	Covariance	75	
3.1.2	Correlation	76	
3.1.3	Uncorrelated does not imply independent	77	
3.1.4	Correlation does not imply causation	77	
3.1.5	Simpson's paradox	78	
3.2	The multivariate Gaussian (normal) distribution	79	
3.2.1	Definition	79	
3.2.2	Mahalanobis distance	81	
3.2.3	Marginals and conditionals of an MVN *	82	
3.2.4	Example: conditioning a 2d Gaussian	83	
3.2.5	Example: Imputing missing values *	83	
3.3	Linear Gaussian systems *	84	
3.3.1	Bayes rule for Gaussians	85	
3.3.2	Derivation *	85	
3.3.3	Example: Inferring an unknown scalar	86	
3.3.4	Example: inferring an unknown vector	88	
3.3.5	Example: sensor fusion	89	
3.4	The exponential family *	90	
3.4.1	Definition	90	
3.4.2	Example	91	
3.4.3	Log partition function is cumulant generating function	92	
3.4.4	Maximum entropy derivation of the exponential family	92	
3.5	Mixture models	93	
3.5.1	Gaussian mixture models	94	
3.5.2	Bernoulli mixture models	95	
3.6	Probabilistic graphical models *	96	
3.6.1	Representation	97	
3.6.2	Inference	99	
3.6.3	Learning	100	
3.7	Exercises	100	
4	Statistics	103	
4.1	Introduction	103	
4.2	Maximum likelihood estimation (MLE)	103	
4.2.1	Definition	103	
4.2.2	Justification for MLE	104	
4.2.3	Example: MLE for the Bernoulli distribution	106	
4.2.4	Example: MLE for the categorical distribution	107	
4.2.5	Example: MLE for the univariate Gaussian	107	
4.2.6	Example: MLE for the multivariate Gaussian	108	
4.2.7	Example: MLE for linear regression	110	
4.3	Empirical risk minimization (ERM)	111	
4.3.1	Example: minimizing the misclassification rate	111	

4.3.2	Surrogate loss	112	
4.4	Other estimation methods *	112	
4.4.1	The method of moments	112	
4.4.2	Online (recursive) estimation	114	
4.5	Regularization	116	
4.5.1	Example: MAP estimation for the Bernoulli distribution	117	
4.5.2	Example: MAP estimation for the multivariate Gaussian *	118	
4.5.3	Example: weight decay	119	
4.5.4	Picking the regularizer using a validation set	120	
4.5.5	Cross-validation	121	
4.5.6	Early stopping	123	
4.5.7	Using more data	123	
4.6	Bayesian statistics *	124	
4.6.1	Conjugate priors	125	
4.6.2	The beta-binomial model	125	
4.6.3	The Dirichlet-multinomial model	133	
4.6.4	The Gaussian-Gaussian model	137	
4.6.5	Beyond conjugate priors	140	
4.6.6	Credible intervals	141	
4.6.7	Bayesian machine learning	143	
4.6.8	Computational issues	147	
4.7	Frequentist statistics *	150	
4.7.1	Sampling distributions	150	
4.7.2	Gaussian approximation of the sampling distribution of the MLE	151	
4.7.3	Bootstrap approximation of the sampling distribution of any estimator	151	
4.7.4	Confidence intervals	153	
4.7.5	Caution: Confidence intervals are not credible	154	
4.7.6	The bias-variance tradeoff	155	
4.8	Exercises	160	
5	Decision Theory	163	
5.1	Bayesian decision theory	163	
5.1.1	Basics	163	
5.1.2	Classification problems	165	
5.1.3	ROC curves	167	
5.1.4	Precision-recall curves	170	
5.1.5	Regression problems	172	
5.1.6	Probabilistic prediction problems	173	
5.2	Bayesian hypothesis testing	175	
5.2.1	Example: Testing if a coin is fair	176	
5.2.2	Bayesian model selection	177	
5.2.3	Occam's razor	178	
5.2.4	Connection between cross validation and marginal likelihood	179	
5.2.5	Information criteria	180	
5.3	Frequentist decision theory	182	

5.3.1	Computing the risk of an estimator	182
5.3.2	Consistent estimators	185
5.3.3	Admissible estimators	185
5.4	Empirical risk minimization	186
5.4.1	Empirical risk	186
5.4.2	Structural risk	188
5.4.3	Cross-validation	189
5.4.4	Statistical learning theory *	189
5.5	Frequentist hypothesis testing *	191
5.5.1	Likelihood ratio test	191
5.5.2	Null hypothesis significance testing (NHST)	192
5.5.3	p-values	193
5.5.4	p-values considered harmful	193
5.5.5	Why isn't everyone a Bayesian?	195
5.6	Exercises	197
6	Information Theory	199
6.1	Entropy	199
6.1.1	Entropy for discrete random variables	199
6.1.2	Cross entropy	201
6.1.3	Joint entropy	201
6.1.4	Conditional entropy	202
6.1.5	Perplexity	203
6.1.6	Differential entropy for continuous random variables *	204
6.2	Relative entropy (KL divergence) *	205
6.2.1	Definition	205
6.2.2	Interpretation	206
6.2.3	Example: KL divergence between two Gaussians	206
6.2.4	Non-negativity of KL	206
6.2.5	KL divergence and MLE	207
6.2.6	Forward vs reverse KL	208
6.3	Mutual information *	209
6.3.1	Definition	209
6.3.2	Interpretation	210
6.3.3	Example	210
6.3.4	Conditional mutual information	211
6.3.5	MI as a “generalized correlation coefficient”	212
6.3.6	Normalized mutual information	213
6.3.7	Maximal information coefficient	213
6.3.8	Data processing inequality	215
6.3.9	Sufficient Statistics	216
6.3.10	Fano's inequality *	217
6.4	Exercises	218
7	Linear Algebra	221

7.1	Introduction	221	
7.1.1	Notation	221	
7.1.2	Vector spaces	224	
7.1.3	Norms of a vector and matrix	226	
7.1.4	Properties of a matrix	228	
7.1.5	Special types of matrices	231	
7.2	Matrix multiplication	234	
7.2.1	Vector–vector products	234	
7.2.2	Matrix–vector products	235	
7.2.3	Matrix–matrix products	235	
7.2.4	Application: manipulating data matrices	237	
7.2.5	Kronecker products *	240	
7.2.6	Einstein summation *	240	
7.3	Matrix inversion	241	
7.3.1	The inverse of a square matrix	241	
7.3.2	Schur complements *	242	
7.3.3	The matrix inversion lemma *	243	
7.3.4	Matrix determinant lemma *	243	
7.3.5	Application: deriving the conditionals of an MVN *	244	
7.4	Eigenvalue decomposition (EVD)	245	
7.4.1	Basics	245	
7.4.2	Diagonalization	246	
7.4.3	Eigenvalues and eigenvectors of symmetric matrices	247	
7.4.4	Geometry of quadratic forms	248	
7.4.5	Standardizing and whitening data	248	
7.4.6	Power method	250	
7.4.7	Deflation	251	
7.4.8	Eigenvectors optimize quadratic forms	251	
7.5	Singular value decomposition (SVD)	251	
7.5.1	Basics	251	
7.5.2	Connection between SVD and EVD	252	
7.5.3	Pseudo inverse	253	
7.5.4	SVD and the range and null space of a matrix *	254	
7.5.5	Truncated SVD	256	
7.6	Other matrix decompositions *	256	
7.6.1	LU factorization	256	
7.6.2	QR decomposition	257	
7.6.3	Cholesky decomposition	258	
7.7	Solving systems of linear equations *	258	
7.7.1	Solving square systems	259	
7.7.2	Solving underconstrained systems (least norm estimation)	259	
7.7.3	Solving overconstrained systems (least squares estimation)	261	
7.8	Matrix calculus	261	
7.8.1	Derivatives	262	
7.8.2	Gradients	262	

7.8.3	Directional derivative	263
7.8.4	Total derivative *	263
7.8.5	Jacobian	263
7.8.6	Hessian	264
7.8.7	Gradients of commonly used functions	265
7.9	Exercises	266
8	Optimization	269
8.1	Introduction	269
8.1.1	Local vs global optimization	269
8.1.2	Constrained vs unconstrained optimization	271
8.1.3	Convex vs nonconvex optimization	271
8.1.4	Smooth vs nonsmooth optimization	275
8.2	First-order methods	276
8.2.1	Descent direction	278
8.2.2	Step size (learning rate)	278
8.2.3	Convergence rates	280
8.2.4	Momentum methods	281
8.3	Second-order methods	283
8.3.1	Newton's method	283
8.3.2	BFGS and other quasi-Newton methods	284
8.3.3	Trust region methods	285
8.4	Stochastic gradient descent	286
8.4.1	Application to finite sum problems	287
8.4.2	Example: SGD for fitting linear regression	287
8.4.3	Choosing the step size (learning rate)	288
8.4.4	Iterate averaging	291
8.4.5	Variance reduction *	291
8.4.6	Preconditioned SGD	292
8.5	Constrained optimization	295
8.5.1	Lagrange multipliers	296
8.5.2	The KKT conditions	297
8.5.3	Linear programming	299
8.5.4	Quadratic programming	300
8.5.5	Mixed integer linear programming *	301
8.6	Proximal gradient method *	301
8.6.1	Projected gradient descent	302
8.6.2	Proximal operator for ℓ_1 -norm regularizer	303
8.6.3	Proximal operator for quantization	304
8.6.4	Incremental (online) proximal methods	305
8.7	Bound optimization *	306
8.7.1	The general algorithm	306
8.7.2	The EM algorithm	306
8.7.3	Example: EM for a GMM	309
8.8	Blackbox and derivative free optimization	313

8.9	Exercises	314
II	Linear Models	315
9	Linear Discriminant Analysis	317
9.1	Introduction	317
9.2	Gaussian discriminant analysis	317
9.2.1	Quadratic decision boundaries	318
9.2.2	Linear decision boundaries	319
9.2.3	The connection between LDA and logistic regression	319
9.2.4	Model fitting	320
9.2.5	Nearest centroid classifier	322
9.2.6	Fisher's linear discriminant analysis *	322
9.3	Naive Bayes classifiers	326
9.3.1	Example models	326
9.3.2	Model fitting	327
9.3.3	Bayesian naive Bayes	328
9.3.4	The connection between naive Bayes and logistic regression	329
9.4	Generative vs discriminative classifiers	330
9.4.1	Advantages of discriminative classifiers	330
9.4.2	Advantages of generative classifiers	331
9.4.3	Handling missing features	331
9.5	Exercises	332
10	Logistic Regression	333
10.1	Introduction	333
10.2	Binary logistic regression	333
10.2.1	Linear classifiers	333
10.2.2	Nonlinear classifiers	334
10.2.3	Maximum likelihood estimation	336
10.2.4	Stochastic gradient descent	339
10.2.5	Perceptron algorithm	340
10.2.6	Iteratively reweighted least squares	340
10.2.7	MAP estimation	342
10.2.8	Standardization	343
10.3	Multinomial logistic regression	344
10.3.1	Linear and nonlinear classifiers	345
10.3.2	Maximum likelihood estimation	345
10.3.3	Gradient-based optimization	347
10.3.4	Bound optimization	347
10.3.5	MAP estimation	349
10.3.6	Maximum entropy classifiers	350
10.3.7	Hierarchical classification	351
10.3.8	Handling large numbers of classes	352

10.4	Robust logistic regression *	353	
10.4.1	Mixture model for the likelihood	353	
10.4.2	Bi-tempered loss	354	
10.5	Bayesian logistic regression *	357	
10.5.1	Laplace approximation	357	
10.5.2	Approximating the posterior predictive	358	
10.6	Exercises	361	
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	377	
11.3.3	Choosing the strength of the regularizer	378	
11.4	Lasso regression	379	
11.4.1	MAP estimation with a Laplace prior (ℓ_1 regularization)	379	
11.4.2	Why does ℓ_1 regularization yield sparse solutions?	380	
11.4.3	Hard vs soft thresholding	381	
11.4.4	Regularization path	383	
11.4.5	Comparison of least squares, lasso, ridge and subset selection	384	
11.4.6	Variable selection consistency	386	
11.4.7	Group lasso	387	
11.4.8	Elastic net (ridge and lasso combined)	390	
11.4.9	Optimization algorithms	391	
11.5	Regression splines *	393	
11.5.1	B-spline basis functions	393	
11.5.2	Fitting a linear model using a spline basis	395	
11.5.3	Smoothing splines	395	
11.5.4	Generalized additive models	395	
11.6	Robust linear regression *	396	
11.6.1	Laplace likelihood	396	
11.6.2	Student- t likelihood	398	
11.6.3	Huber loss	398	
11.6.4	RANSAC	398	
11.7	Bayesian linear regression *	399	
11.7.1	Priors	399	
11.7.2	Posteriors	399	
11.7.3	Example	400	
11.7.4	Computing the posterior predictive	400	
11.7.5	The advantage of centering	402	

11.7.6	Dealing with multicollinearity	403	
11.7.7	Automatic relevancy determination (ARD) *	404	
11.8	Exercises	405	
12	Generalized Linear Models *	409	
12.1	Introduction	409	
12.2	Examples	409	
12.2.1	Linear regression	410	
12.2.2	Binomial regression	410	
12.2.3	Poisson regression	411	
12.3	GLMs with non-canonical link functions	411	
12.4	Maximum likelihood estimation	412	
12.5	Worked example: predicting insurance claims	413	
III	Deep Neural Networks	417	
13	Neural Networks for Structured Data	419	
13.1	Introduction	419	
13.2	Multilayer perceptrons (MLPs)	420	
13.2.1	The XOR problem	421	
13.2.2	Differentiable MLPs	422	
13.2.3	Activation functions	422	
13.2.4	Example models	423	
13.2.5	The importance of depth	428	
13.2.6	The “deep learning revolution”	429	
13.2.7	Connections with biology	429	
13.3	Backpropagation	432	
13.3.1	Forward vs reverse mode differentiation	432	
13.3.2	Reverse mode differentiation for multilayer perceptrons	434	
13.3.3	Vector-Jacobian product for common layers	436	
13.3.4	Computation graphs	438	
13.4	Training neural networks	440	
13.4.1	Tuning the learning rate	441	
13.4.2	Vanishing and exploding gradients	441	
13.4.3	Non-saturating activation functions	442	
13.4.4	Residual connections	445	
13.4.5	Parameter initialization	446	
13.4.6	Parallel training	447	
13.5	Regularization	448	
13.5.1	Early stopping	448	
13.5.2	Weight decay	449	
13.5.3	Sparse DNNs	449	
13.5.4	Dropout	449	
13.5.5	Bayesian neural networks	451	

13.5.6	Regularization effects of (stochastic) gradient descent *	451
13.6	Other kinds of feedforward networks *	453
13.6.1	Radial basis function networks	453
13.6.2	Mixtures of experts	454
13.7	Exercises	457
14	Neural Networks for Images	461
14.1	Introduction	461
14.2	Common layers	462
14.2.1	Convolutional layers	462
14.2.2	Pooling layers	469
14.2.3	Putting it all together	470
14.2.4	Normalization layers	470
14.3	Common architectures for image classification	473
14.3.1	LeNet	473
14.3.2	AlexNet	475
14.3.3	GoogLeNet (Inception)	476
14.3.4	ResNet	477
14.3.5	DenseNet	478
14.3.6	Neural architecture search	479
14.4	Other forms of convolution *	479
14.4.1	Dilated convolution	479
14.4.2	Transposed convolution	481
14.4.3	Depthwise separable convolution	482
14.5	Solving other discriminative vision tasks with CNNs *	482
14.5.1	Image tagging	483
14.5.2	Object detection	483
14.5.3	Instance segmentation	484
14.5.4	Semantic segmentation	484
14.5.5	Human pose estimation	486
14.6	Generating images by inverting CNNs *	487
14.6.1	Converting a trained classifier into a generative model	487
14.6.2	Image priors	488
14.6.3	Visualizing the features learned by a CNN	490
14.6.4	Deep Dream	490
14.6.5	Neural style transfer	491
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)	500
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	513	
15.4.2	Kernel regression as non-parametric attention	514	
15.4.3	Parametric attention	514	
15.4.4	Seq2Seq with attention	515	
15.4.5	Seq2vec with attention (text classification)	518	
15.4.6	Seq+Seq2Vec with attention (text pair classification)	518	
15.4.7	Soft vs hard attention	519	
15.5	Transformers	520	
15.5.1	Self-attention	520	
15.5.2	Multi-headed attention	521	
15.5.3	Positional encoding	522	
15.5.4	Putting it all together	523	
15.5.5	Comparing transformers, CNNs and RNNs	525	
15.5.6	Transformers for images *	526	
15.5.7	Other transformer variants *	526	
15.6	Efficient transformers *	527	
15.6.1	Fixed non-learnable localized attention patterns	527	
15.6.2	Learnable sparse attention patterns	528	
15.6.3	Memory and recurrence methods	529	
15.6.4	Low-rank and kernel methods	529	
15.7	Language models and unsupervised representation learning	531	
15.7.1	ELMo	531	
15.7.2	BERT	532	
15.7.3	GPT	536	
15.7.4	T5	536	
15.7.5	Discussion	537	

IV Nonparametric Models 539

16 Exemplar-based Methods 541

16.1	K nearest neighbor (KNN) classification	541	
16.1.1	Example	542	
16.1.2	The curse of dimensionality	542	
16.1.3	Reducing the speed and memory requirements	544	
16.1.4	Open set recognition	544	
16.2	Learning distance metrics	545	
16.2.1	Linear and convex methods	546	

16.2.2	Deep metric learning	548	
16.2.3	Classification losses	548	
16.2.4	Ranking losses	549	
16.2.5	Speeding up ranking loss optimization		550
16.2.6	Other training tricks for DML	553	
16.3	Kernel density estimation (KDE)	554	
16.3.1	Density kernels	554	
16.3.2	Parzen window density estimator		555
16.3.3	How to choose the bandwidth parameter		556
16.3.4	From KDE to KNN classification	557	
16.3.5	Kernel regression	557	
17	Kernel Methods *	561	
17.1	Mercer kernels	561	
17.1.1	Mercer's theorem	562	
17.1.2	Some popular Mercer kernels		563
17.2	Gaussian processes	568	
17.2.1	Noise-free observations	568	
17.2.2	Noisy observations	569	
17.2.3	Comparison to kernel regression		570
17.2.4	Weight space vs function space		571
17.2.5	Numerical issues	571	
17.2.6	Estimating the kernel	572	
17.2.7	GPs for classification	575	
17.2.8	Connections with deep learning		576
17.2.9	Scaling GPs to large datasets		577
17.3	Support vector machines (SVMs)	579	
17.3.1	Large margin classifiers	579	
17.3.2	The dual problem	581	
17.3.3	Soft margin classifiers	583	
17.3.4	The kernel trick	584	
17.3.5	Converting SVM outputs into probabilities		585
17.3.6	Connection with logistic regression		585
17.3.7	Multi-class classification with SVMs		586
17.3.8	How to choose the regularizer C		587
17.3.9	Kernel ridge regression	588	
17.3.10	SVMs for regression	589	
17.4	Sparse vector machines	591	
17.4.1	Relevance vector machines (RVMs)		592
17.4.2	Comparison of sparse and dense kernel methods		592
17.5	Exercises	595	
18	Trees, Forests, Bagging, and Boosting	597	
18.1	Classification and regression trees (CART)	597	
18.1.1	Model definition	597	

18.1.2	Model fitting	599	
18.1.3	Regularization	600	
18.1.4	Handling missing input features	600	
18.1.5	Pros and cons	600	
18.2	Ensemble learning	602	
18.2.1	Stacking	602	
18.2.2	Ensembling is not Bayes model averaging	603	
18.3	Bagging	603	
18.4	Random forests	604	
18.5	Boosting	605	
18.5.1	Forward stagewise additive modeling	606	
18.5.2	Quadratic loss and least squares boosting	606	
18.5.3	Exponential loss and AdaBoost	607	
18.5.4	LogitBoost	610	
18.5.5	Gradient boosting	610	
18.6	Interpreting tree ensembles	614	
18.6.1	Feature importance	615	
18.6.2	Partial dependency plots	617	

V Beyond Supervised Learning 619

19	Learning with Fewer Labeled Examples	621	
19.1	Data augmentation	621	
19.1.1	Examples	621	
19.1.2	Theoretical justification	622	
19.2	Transfer learning	622	
19.2.1	Fine-tuning	623	
19.2.2	Adapters	624	
19.2.3	Supervised pre-training	625	
19.2.4	Unsupervised pre-training (self-supervised learning)	626	
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	633	
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 *	640	
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	644	
19.4.3	Batch active learning	645	
19.5	Meta-learning	645	

19.5.1	Model-agnostic meta-learning (MAML)	646
19.6	Few-shot learning	647
19.6.1	Matching networks	648
19.7	Weakly supervised learning	649
19.8	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	670
20.3	Autoencoders	673
20.3.1	Bottleneck autoencoders	674
20.3.2	Denoising autoencoders	676
20.3.3	Contractive autoencoders	676
20.3.4	Sparse autoencoders	677
20.3.5	Variational autoencoders	677
20.4	Manifold learning *	682
20.4.1	What are manifolds?	683
20.4.2	The manifold hypothesis	683
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	689
20.4.7	Maximum variance unfolding (MVU)	691
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	
21.2.3	Extensions	715	
21.3	K means clustering	716	
21.3.1	The algorithm	716	
21.3.2	Examples	716	
21.3.3	Vector quantization	718	
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	727	
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	731	
21.6	Biclustering *	731	
21.6.1	Basic biclustering	732	
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	741	
22.2.1	Bayesian personalized ranking	741	
22.2.2	Factorization machines	742	
22.2.3	Neural matrix factorization	743	
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	751	
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	755	
23.4	Graph Neural Networks	756	
23.4.1	Message passing GNNs	756	
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	760	
23.5.2	Semi-supervised embeddings	762	
23.6	Applications	763	
23.6.1	Unsupervised applications	763	
23.6.2	Supervised applications	765	
A	Notation	767	
A.1	Introduction	767	
A.2	Common mathematical symbols	767	
A.3	Functions	768	
A.3.1	Common functions of one argument	768	
A.3.2	Common functions of two arguments	768	
A.3.3	Common functions of > 2 arguments	768	
A.4	Linear algebra	769	
A.4.1	General notation	769	
A.4.2	Vectors	769	
A.4.3	Matrices	769	
A.4.4	Matrix calculus	770	
A.5	Optimization	770	
A.6	Probability	771	
A.7	Information theory	771	
A.8	Statistics and machine learning	772	
A.8.1	Supervised learning	772	
A.8.2	Unsupervised learning and generative models	772	
A.8.3	Bayesian inference	772	
A.9	Abbreviations	773	
Index	775		
Bibliography	792		