

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”	15
1.3.3	Self-supervised learning	16
1.3.4	Evaluating unsupervised learning	16
1.4	Reinforcement learning	17
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	22
1.5.4	Preprocessing text data	23
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	27
1.6.3	Caveats	27

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	35	
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	46	
2.3.3	Inverse problems *	47	
2.4	Bernoulli and binomial distributions	48	
2.4.1	Definition	48	
2.4.2	Sigmoid (logistic) function	49	
2.4.3	Binary logistic regression	51	
2.5	Categorical and multinomial distributions	51	
2.5.1	Definition	52	
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	57	
2.6.3	Regression	58	
2.6.4	Why is the Gaussian distribution so widely used?	59	
2.6.5	Dirac delta function as a limiting case	59	
2.7	Some other common univariate distributions *	60	
2.7.1	Student t distribution	60	
2.7.2	Cauchy distribution	61	
2.7.3	Laplace distribution	62	
2.7.4	Beta distribution	62	
2.7.5	Gamma distribution	63	
2.7.6	Empirical distribution	64	
2.8	Transformations of random variables *	65	
2.8.1	Discrete case	65	
2.8.2	Continuous case	65	
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	70	
2.8.7	Monte Carlo approximation	71	

2.9	Exercises	72	
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	Simpsons' paradox	78	
3.2	The multivariate Gaussian (normal) distribution	78	
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: Imputing missing values *	83	
3.3	Linear Gaussian systems *	83	
3.3.1	Example: inferring a latent vector from a noisy sensor	85	
3.3.2	Example: inferring a latent vector from multiple noisy sensors	86	
3.4	The exponential family	86	
3.4.1	Definition	86	
3.4.2	Example	87	
3.4.3	Log partition function is cumulant generating function	88	
3.4.4	Maximum entropy derivation of the exponential family	88	
3.5	Mixture models	89	
3.5.1	Gaussian mixture models	90	
3.5.2	Mixtures of Bernoullis	91	
3.6	Probabilistic graphical models *	92	
3.6.1	Representation	93	
3.6.2	Inference	95	
3.6.3	Learning	96	
3.7	Exercises	96	
4	Statistics	101	
4.1	Introduction	101	
4.2	Maximum likelihood estimation (MLE)	101	
4.2.1	Definition	101	
4.2.2	Justification for MLE	102	
4.2.3	Example: MLE for the Bernoulli distribution	104	
4.2.4	Example: MLE for the categorical distribution	105	
4.2.5	Example: MLE for the univariate Gaussian	105	
4.2.6	Example: MLE for the multivariate Gaussian	106	
4.2.7	Example: MLE for linear regression	108	
4.3	Empirical risk minimization (ERM)	109	
4.3.1	Example: minimizing the misclassification rate	109	
4.3.2	Surrogate loss	110	
4.4	Other estimation methods *	110	

4.4.1	The method of moments	110	
4.4.2	Online (recursive) estimation	112	
4.5	Regularization	114	
4.5.1	Example: MAP estimation for the Bernoulli distribution	115	
4.5.2	Example: MAP estimation for the multivariate Gaussian *	116	
4.5.3	Example: weight decay	117	
4.5.4	Picking the regularizer using a validation set	118	
4.5.5	Cross-validation	119	
4.5.6	Early stopping	120	
4.5.7	Using more data	121	
4.6	Bayesian statistics *	121	
4.6.1	Conjugate priors	123	
4.6.2	The beta-binomial model	123	
4.6.3	The Dirichlet-multinomial model	131	
4.6.4	The Gaussian-Gaussian model	134	
4.6.5	Beyond conjugate priors	137	
4.6.6	Credible intervals	139	
4.6.7	Bayesian machine learning	140	
4.6.8	Computational issues	145	
4.7	Frequentist statistics *	147	
4.7.1	Sampling distributions	148	
4.7.2	Gaussian approximation of the sampling distribution of the MLE	148	
4.7.3	Bootstrap approximation of the sampling distribution of any estimator	149	
4.7.4	Confidence intervals	150	
4.7.5	Caution: Confidence intervals are not credible	151	
4.7.6	The bias-variance tradeoff	152	
4.8	Exercises	157	
5	Decision Theory	161	
5.1	Bayesian decision theory	161	
5.1.1	Basics	161	
5.1.2	Classification problems	163	
5.1.3	ROC curves	165	
5.1.4	Precision-recall curves	167	
5.1.5	Regression problems	170	
5.1.6	Probabilistic prediction problems	171	
5.2	Bayesian hypothesis testing	173	
5.2.1	Example: Testing if a coin is fair	174	
5.2.2	Bayesian model selection	174	
5.2.3	Occam's razor	176	
5.2.4	Connection between cross validation and marginal likelihood	178	
5.2.5	Information criteria	179	
5.3	Frequentist decision theory	180	
5.3.1	Computing the risk of an estimator	180	
5.3.2	Consistent estimators	183	

5.3.3	Admissible estimators	183	
5.4	Empirical risk minimization	184	
5.4.1	Empirical risk	184	
5.4.2	Structural risk	186	
5.4.3	Cross-validation	187	
5.4.4	Statistical learning theory *	187	
5.5	Frequentist hypothesis testing *	189	
5.5.1	Likelihood ratio test	189	
5.5.2	Null hypothesis significance testing (NHST)	190	
5.5.3	p-values	191	
5.5.4	p-values considered harmful	191	
5.5.5	Why isn't everyone a Bayesian?	193	
5.6	Exercises	195	
6	Information Theory	197	
6.1	Entropy	197	
6.1.1	Entropy for discrete random variables	197	
6.1.2	Cross entropy	199	
6.1.3	Joint entropy	199	
6.1.4	Conditional entropy	200	
6.1.5	Perplexity	201	
6.1.6	Differential entropy for continuous random variables *	202	
6.2	Relative entropy (KL divergence) *	203	
6.2.1	Definition	203	
6.2.2	Interpretation	204	
6.2.3	Example: KL divergence between two Gaussians	204	
6.2.4	Non-negativity of KL	204	
6.2.5	KL divergence and MLE	205	
6.2.6	Forward vs reverse KL	206	
6.3	Mutual information *	207	
6.3.1	Definition	207	
6.3.2	Interpretation	208	
6.3.3	Example	208	
6.3.4	Conditional mutual information	209	
6.3.5	MI as a “generalized correlation coefficient”	210	
6.3.6	Normalized mutual information	211	
6.3.7	Maximal information coefficient	211	
6.3.8	Data processing inequality	213	
6.3.9	Sufficient Statistics	214	
6.3.10	Fano's inequality *	215	
6.4	Exercises	216	
7	Linear Algebra	219	
7.1	Introduction	219	
7.1.1	Notation	219	

7.1.2	Vector spaces	222	
7.1.3	Norms of a vector and matrix	224	
7.1.4	Properties of a matrix	226	
7.1.5	Special types of matrices	229	
7.2	Matrix multiplication	232	
7.2.1	Vector-Vector Products	232	
7.2.2	Matrix-Vector Products	233	
7.2.3	Matrix-Matrix Products	233	
7.2.4	Application: manipulating data matrices	235	
7.2.5	Kronecker products *	237	
7.2.6	Einstein summation *	238	
7.3	Matrix inversion	239	
7.3.1	The inverse of a square matrix	239	
7.3.2	Schur complements *	239	
7.3.3	The matrix inversion lemma *	241	
7.3.4	Matrix determinant lemma *	241	
7.4	Eigenvalue decomposition (EVD)	242	
7.4.1	Basics	242	
7.4.2	Diagonalization	243	
7.4.3	Eigenvalues and eigenvectors of symmetric matrices	243	
7.4.4	Geometry of quadratic forms	244	
7.4.5	Standardizing and whitening data	244	
7.4.6	Power method	245	
7.4.7	Deflation	247	
7.4.8	Eigenvectors optimize quadratic forms	247	
7.5	Singular value decomposition (SVD)	248	
7.5.1	Basics	248	
7.5.2	Connection between SVD and EVD	249	
7.5.3	Pseudo inverse	249	
7.5.4	SVD and the range and null space of a matrix *	250	
7.5.5	Truncated SVD	251	
7.6	Other matrix decompositions *	252	
7.6.1	LU factorization	252	
7.6.2	QR decomposition	253	
7.6.3	Cholesky decomposition	254	
7.7	Solving systems of linear equations *	254	
7.7.1	Solving square systems	255	
7.7.2	Solving underconstrained systems (least norm estimation)	256	
7.7.3	Solving overconstrained systems (least squares estimation)	257	
7.8	Matrix calculus	258	
7.8.1	Derivatives	258	
7.8.2	Gradients	259	
7.8.3	Directional derivative	259	
7.8.4	Total derivative *	260	
7.8.5	Jacobian	260	

7.8.6	Hessian	261
7.8.7	Gradients of commonly used functions	261
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	286
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	290
8.4.5	Variance reduction *	290
8.4.6	Preconditioned SGD	291
8.5	Constrained optimization	294
8.5.1	Lagrange multipliers	295
8.5.2	The KKT conditions	296
8.5.3	Linear programming	298
8.5.4	Quadratic programming	299
8.5.5	Mixed integer linear programming *	300
8.6	Proximal gradient method *	300
8.6.1	Projected gradient descent	301
8.6.2	Proximal operator for ℓ_1 -norm regularizer	302
8.6.3	Proximal operator for quantization	303
8.7	Bound optimization *	304
8.7.1	The general algorithm	305
8.7.2	The EM algorithm	305
8.7.3	Example: EM for a GMM	308
8.8	Blackbox and derivative free optimization	312
8.9	Exercises	313

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	339
10.2.6	Iteratively reweighted least squares	340
10.2.7	MAP estimation	341
10.2.8	Standardization	343
10.3	Multinomial logistic regression	344
10.3.1	Linear and nonlinear classifiers	344
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	363
11.1	Introduction	363
11.2	Least squares linear regression	363
11.2.1	Terminology	363
11.2.2	Least squares estimation	364
11.2.3	Other approaches to computing the MLE	368
11.2.4	Measuring goodness of fit	372
11.3	Ridge regression	373
11.3.1	Computing the MAP estimate	374
11.3.2	Connection between ridge regression and PCA	376
11.3.3	Choosing the strength of the regularizer	377
11.4	Lasso regression	377
11.4.1	MAP estimation with a Laplace prior (ℓ_1 regularization)	378
11.4.2	Why does ℓ_1 regularization yield sparse solutions?	379
11.4.3	Hard vs soft thresholding	380
11.4.4	Regularization path	381
11.4.5	Comparison of least squares, lasso, ridge and subset selection	383
11.4.6	Variable selection consistency	384
11.4.7	Group lasso	386
11.4.8	Elastic net (ridge and lasso combined)	387
11.4.9	Optimization algorithms	389
11.5	Regression splines *	391
11.5.1	B-spline basis functions	392
11.5.2	Fitting a linear model using a spline basis	393
11.5.3	Smoothing splines	394
11.5.4	Generalized additive models	394
11.6	Robust linear regression *	394
11.6.1	Laplace likelihood	394
11.6.2	Student- t likelihood	396
11.6.3	Huber loss	396
11.6.4	RANSAC	397
11.7	Bayesian linear regression *	397
11.7.1	Priors	397
11.7.2	Posteriors	397
11.7.3	Example	398
11.7.4	Computing the posterior predictive	399
11.7.5	The advantage of centering	400
11.7.6	Dealing with multicollinearity	401
11.8	Exercises	403

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	
12.5 Worked example: predicting insurance claims	409	
 III Deep neural networks	 413	
13 Neural Networks for Structured Data	415	
13.1 Introduction	415	
13.2 Multilayer perceptrons (MLPs)	416	
13.2.1 The XOR problem	417	
13.2.2 Differentiable MLPs	418	
13.2.3 Activation functions	418	
13.2.4 Example models	419	
13.2.5 The importance of depth	424	
13.2.6 The “deep learning revolution”	424	
13.2.7 Connections with biology	425	
13.3 Backpropagation	427	
13.3.1 Forward vs reverse mode differentiation	428	
13.3.2 Reverse mode differentiation for multilayer perceptrons	430	
13.3.3 Vector-Jacobian product for common layers	431	
13.3.4 Computation graphs	434	
13.4 Training neural networks	436	
13.4.1 Tuning the learning rate	436	
13.4.2 Vanishing and exploding gradients	436	
13.4.3 Non-saturating activation functions	437	
13.4.4 Residual connections	440	
13.4.5 Parameter initialization	442	
13.4.6 Parallel training	444	
13.5 Regularization	445	
13.5.1 Early stopping	446	
13.5.2 Weight decay	446	
13.5.3 Sparse DNNs	446	
13.5.4 Dropout	447	
13.5.5 Bayesian neural networks	448	
13.5.6 Regularization effects of (stochastic) gradient descent *	448	
13.6 Other kinds of feedforward networks	450	
13.6.1 Radial basis function networks	450	
13.6.2 Mixtures of experts	451	

13.7	Exercises	454	
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	471	
14.3.5	DenseNet	474	
14.3.6	Neural architecture search	474	
14.4	Other forms of convolution *	475	
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	478	
14.5.1	Image tagging	478	
14.5.2	Object detection	478	
14.5.3	Instance segmentation	479	
14.5.4	Semantic segmentation	480	
14.5.5	Human pose estimation	481	
14.6	Generating images by inverting CNNs *	482	
14.6.1	Converting a trained classifier into a generative model	482	
14.6.2	Image priors	483	
14.6.3	Visualizing the features learned by a CNN	484	
14.6.4	Deep Dream	485	
14.6.5	Neural style transfer	486	
15	Neural networks for sequences	491	
15.1	Introduction	491	
15.2	Recurrent neural networks (RNNs)	491	
15.2.1	Vec2Seq (sequence generation)	491	
15.2.2	Seq2Vec (sequence classification)	494	
15.2.3	Seq2Seq (sequence translation)	495	
15.2.4	Teacher forcing	497	
15.2.5	Backpropagation through time	498	
15.2.6	Vanishing and exploding gradients	499	
15.2.7	Gating and long term memory	500	
15.2.8	Beam search	503	
15.3	1d CNNs	504	

15.3.1	1d CNNs for sequence classification	504	
15.3.2	Causal 1d CNNs for sequence generation	505	
15.4	Attention	506	
15.4.1	Attention as soft dictionary lookup	506	
15.4.2	Kernel regression as non-parametric attention	507	
15.4.3	Parametric attention	508	
15.4.4	Seq2Seq with attention	509	
15.4.5	Seq2vec with attention (text classification)	511	
15.4.6	Seq+Seq2Vec with attention (text pair classification)	511	
15.4.7	Soft vs hard attention	513	
15.5	Transformers	514	
15.5.1	Self-attention	514	
15.5.2	Multi-headed attention	514	
15.5.3	Positional encoding	515	
15.5.4	Putting it altogether	517	
15.5.5	Comparing transformers, CNNs and RNNs	518	
15.5.6	Transformers for images *	519	
15.5.7	Other transformer variants *	520	
15.6	Efficient transformers *	520	
15.6.1	Fixed non-learnable localized attention patterns	521	
15.6.2	Learnable sparse attention patterns	521	
15.6.3	Memory and recurrence methods	522	
15.6.4	Low-rank and kernel methods	522	
15.7	Language models and unsupervised representation learning	524	
15.7.1	ELMo	524	
15.7.2	BERT	525	
15.7.3	GPT	529	
15.7.4	T5	530	
15.7.5	Discussion	530	

IV Nonparametric models 531

16 Exemplar-based Methods 533

16.1	K nearest neighbor (KNN) classification	533	
16.1.1	Example	534	
16.1.2	The curse of dimensionality	534	
16.1.3	Reducing the speed and memory requirements	536	
16.1.4	Open set recognition	536	
16.2	Learning distance metrics	537	
16.2.1	Linear and convex methods	538	
16.2.2	Deep metric learning	539	
16.2.3	Classification losses	540	
16.2.4	Ranking losses	540	
16.2.5	Speeding up ranking loss optimization	542	

16.2.6	Other training tricks for DML	545	
16.3	Kernel density estimation (KDE)	545	
16.3.1	Density kernels	546	
16.3.2	Parzen window density estimator	546	
16.3.3	How to choose the bandwidth parameter	548	
16.3.4	From KDE to KNN classification	548	
16.3.5	Kernel regression	549	
17	Kernel Methods	553	
17.1	Inferring functions from data	553	
17.1.1	Smoothness prior	554	
17.1.2	Inference from noise-free observations	554	
17.1.3	Inference from noisy observations	556	
17.2	Mercer kernels	556	
17.2.1	Mercer's theorem	557	
17.2.2	Some popular Mercer kernels	557	
17.3	Gaussian processes	562	
17.3.1	Noise-free observations	562	
17.3.2	Noisy observations	563	
17.3.3	Comparison to kernel regression	565	
17.3.4	Weight space vs function space	565	
17.3.5	Numerical issues	566	
17.3.6	Estimating the kernel	566	
17.3.7	GPs for classification	569	
17.3.8	Connections with deep learning	571	
17.4	Scaling GPs to large datasets	571	
17.4.1	Sparse (inducing-point) approximations	571	
17.4.2	Exploiting parallelization and kernel matrix structure	571	
17.4.3	Random feature approximation	572	
17.5	Support vector machines (SVMs)	573	
17.5.1	Large margin classifiers	573	
17.5.2	The dual problem	576	
17.5.3	Soft margin classifiers	577	
17.5.4	The kernel trick	578	
17.5.5	Converting SVM outputs into probabilities	579	
17.5.6	Connection with logistic regression	580	
17.5.7	Multi-class classification with SVMs	580	
17.5.8	How to choose the regularizer C	581	
17.5.9	Kernel ridge regression	583	
17.5.10	SVMs for regression	584	
17.6	Sparse vector machines	585	
17.6.1	Relevance vector machines (RVMs)	587	
17.6.2	Comparison of sparse and dense kernel methods	587	
17.7	Exercises	590	

18	Trees, Forests, Bagging and Boosting	591
18.1	Classification and regression trees (CART)	591
18.1.1	Model definition	591
18.1.2	Model fitting	593
18.1.3	Regularization	594
18.1.4	Handling missing input features	594
18.1.5	Pros and cons	594
18.2	Ensemble learning	596
18.2.1	Stacking	596
18.2.2	Ensembling is not Bayes model averaging	597
18.3	Bagging	597
18.4	Random forests	598
18.5	Boosting	599
18.5.1	Forward stagewise additive modeling	600
18.5.2	Quadratic loss and least squares boosting	600
18.5.3	Exponential loss and AdaBoost	601
18.5.4	LogitBoost	604
18.5.5	Gradient boosting	604
18.6	Interpreting tree ensembles	608
18.6.1	Feature importance	609
18.6.2	Partial dependency plots	610
V	Beyond supervised learning	613
19	Learning with Fewer Labeled Examples	615
19.1	Data augmentation	615
19.1.1	Examples	615
19.1.2	Theoretical justification	616
19.2	Transfer learning	616
19.2.1	Fine-tuning	617
19.2.2	Adapters	618
19.2.3	Supervised pre-training	619
19.2.4	Unsupervised pre-training (self-supervised learning)	620
19.2.5	Domain adaptation	625
19.3	Semi-supervised learning	625
19.3.1	Self-training and pseudo-labeling	626
19.3.2	Entropy minimization	627
19.3.3	Co-training	630
19.3.4	Label propagation on graphs	630
19.3.5	Consistency regularization	631
19.3.6	Deep generative models *	633
19.3.7	Combining self-supervised and semi-supervised learning	636
19.4	Active learning	637
19.4.1	Decision-theoretic approach	638

19.4.2	Information-theoretic approach	638
19.4.3	Batch active learning	639
19.5	Meta-learning	639
19.5.1	Model-agnostic meta-learning (MAML)	639
19.6	Few-shot learning	640
19.6.1	Matching networks	641
19.7	Weakly supervised learning	642
19.8	Exercises	643
20	Dimensionality Reduction	645
20.1	Principal components analysis (PCA)	645
20.1.1	Examples	645
20.1.2	Derivation of the algorithm	647
20.1.3	Computational issues	650
20.1.4	Choosing the number of latent dimensions	652
20.2	Factor analysis *	654
20.2.1	Generative model	655
20.2.2	Probabilistic PCA	656
20.2.3	EM algorithm for FA/PPCA	657
20.2.4	Unidentifiability of the parameters	659
20.2.5	Nonlinear factor analysis	661
20.2.6	Mixtures of factor analysers	662
20.2.7	Exponential family factor analysis	663
20.2.8	Factor analysis models for paired data	665
20.3	Autoencoders	667
20.3.1	Bottleneck autoencoders	668
20.3.2	Denoising autoencoders	669
20.3.3	Contractive autoencoders	670
20.3.4	Sparse autoencoders	671
20.3.5	Variational autoencoders	671
20.4	Manifold learning *	676
20.4.1	What are manifolds?	677
20.4.2	The manifold hypothesis	677
20.4.3	Approaches to manifold learning	678
20.4.4	Multi-dimensional scaling (MDS)	679
20.4.5	Isomap	682
20.4.6	Kernel PCA	683
20.4.7	Maximum variance unfolding (MVU)	685
20.4.8	Local linear embedding (LLE)	685
20.4.9	Laplacian eigenmaps	686
20.4.10	t-SNE	689
20.5	Word embeddings	693
20.5.1	Latent semantic analysis / indexing	693
20.5.2	Word2vec	695
20.5.3	GloVE	697

20.5.4	Word analogies	698	
20.5.5	RAND-WALK model of word embeddings	699	
20.5.6	Contextual word embeddings	699	
20.6	Exercises	700	
21	Clustering	703	
21.1	Introduction	703	
21.1.1	Evaluating the output of clustering methods	703	
21.2	Hierarchical agglomerative clustering	705	
21.2.1	The algorithm	706	
21.2.2	Example	708	
21.2.3	Extensions	709	
21.3	K means clustering	710	
21.3.1	The algorithm	710	
21.3.2	Examples	710	
21.3.3	Vector quantization	712	
21.3.4	The K-means++ algorithm	713	
21.3.5	The K-medoids algorithm	713	
21.3.6	Speedup tricks	714	
21.3.7	Choosing the number of clusters K	715	
21.4	Clustering using mixture models	718	
21.4.1	Mixtures of Gaussians	718	
21.4.2	Mixtures of Bernoullis	722	
21.5	Spectral clustering *	722	
21.5.1	Normalized cuts	722	
21.5.2	Eigenvectors of the graph Laplacian encode the clustering	723	
21.5.3	Example	724	
21.5.4	Connection with other methods	724	
21.6	Biclustering *	725	
21.6.1	Basic biclustering	725	
21.6.2	Nested partition models (Crosscat)	726	
22	Recommender Systems	729	
22.1	Explicit feedback	729	
22.1.1	Datasets	729	
22.1.2	Collaborative filtering	730	
22.1.3	Matrix factorization	731	
22.1.4	Autoencoders	733	
22.2	Implicit feedback	734	
22.2.1	Bayesian personalized ranking	735	
22.2.2	Factorization machines	735	
22.2.3	Neural matrix factorization	736	
22.3	Leveraging side information	737	
22.4	Exploration-exploitation tradeoff	738	
23	Graph Embeddings	741	

23.1	Introduction	741	
23.2	Graph Embedding as an Encoder/Decoder Problem	742	
23.3	Shallow graph embeddings	744	
23.3.1	Unsupervised embeddings	745	
23.3.2	Distance-based: Euclidean methods	745	
23.3.3	Distance-based: non-Euclidean methods	746	
23.3.4	Outer product-based: Matrix factorization methods	746	
23.3.5	Outer product-based: Skip-gram methods	747	
23.3.6	Supervised embeddings	749	
23.4	Graph Neural Networks	750	
23.4.1	Message passing GNNs	750	
23.4.2	Spectral Graph Convolutions	751	
23.4.3	Spatial Graph Convolutions	751	
23.4.4	Non-Euclidean Graph Convolutions	753	
23.5	Deep graph embeddings	753	
23.5.1	Unsupervised embeddings	754	
23.5.2	Semi-supervised embeddings	756	
23.6	Applications	757	
23.6.1	Unsupervised applications	757	
23.6.2	Supervised applications	759	
Appendices		761	
VI Appendix		763	
A Notation		765	
A.1	Introduction	765	
A.2	Common mathematical symbols	765	
A.3	Functions	766	
A.3.1	Common functions of one argument	766	
A.3.2	Common functions of two arguments	766	
A.3.3	Common functions of > 2 arguments	766	
A.4	Linear algebra	767	
A.4.1	General notation	767	
A.4.2	Vectors	767	
A.4.3	Matrices	767	
A.4.4	Matrix calculus	768	
A.5	Optimization	768	
A.6	Probability	769	
A.7	Information theory	769	
A.8	Statistics and machine learning	769	
A.8.1	Supervised learning	770	
A.8.2	Unsupervised learning and generative models	770	
A.8.3	Bayesian inference	770	

A.9 Abbreviations	771
Bibliography	784