

# Contents

<b>Preface</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 What is machine learning?	1
1.2 Supervised learning	1
1.2.1 Classification	2
1.2.2 Regression	7
1.2.3 Overfitting and generalization	12
1.2.4 No free lunch theorem	12
1.3 Unsupervised learning	13
1.3.1 Clustering	14
1.3.2 Self-supervised learning	14
1.3.3 Evaluating unsupervised learning	14
1.4 Reinforcement learning	16
1.5 Discussion	17
1.5.1 The relationship between ML and other fields	17
1.5.2 Structure of the book	17
1.5.3 Caveats	18
<b>I Foundations</b>	<b>19</b>
<b>2 Probability: univariate models</b>	<b>21</b>
2.1 Introduction	21
2.1.1 What is probability?	21
2.1.2 Types of uncertainty	21
2.1.3 Probability as an extension of logic	22
2.2 Random variables	23
2.2.1 Discrete random variables	23
2.2.2 Continuous random variables	24
2.2.3 Sets of related random variables	26
2.2.4 Independence and conditional independence	27

2.2.5	Moments of a distribution	27	
2.3	Bayes' rule	30	
2.3.1	Example: Testing for COVID-19	31	
2.3.2	Example: The Monty Hall problem	32	
2.3.3	Inverse problems *	34	
2.4	Bernoulli and binomial distributions	35	
2.4.1	Definition	35	
2.4.2	Sigmoid (logistic) function	35	
2.4.3	Binary logistic regression	37	
2.5	Categorical and multinomial distributions	38	
2.5.1	Definition	38	
2.5.2	Softmax function	39	
2.5.3	Multiclass logistic regression	40	
2.5.4	Log-sum-exp trick	41	
2.6	Univariate Gaussian (normal) distribution	42	
2.6.1	Cumulative distribution function	42	
2.6.2	Probability density function	43	
2.6.3	Regression	44	
2.6.4	Why is the Gaussian distribution so widely used?	45	
2.6.5	Dirac delta function as a limiting case	46	
2.7	Some other common univariate distributions *	46	
2.7.1	Student $t$ distribution	46	
2.7.2	Cauchy distribution	48	
2.7.3	Laplace distribution	48	
2.7.4	Beta distribution	49	
2.7.5	Gamma distribution	49	
2.7.6	Half-normal	51	
2.8	Transformations of random variables *	51	
2.8.1	Discrete case	51	
2.8.2	Continuous case	51	
2.8.3	Invertible transformations (bijections)	51	
2.8.4	Moments of a linear transformation	54	
2.8.5	The convolution theorem	55	
2.8.6	Central limit theorem	56	
2.8.7	Monte Carlo approximation	57	
2.9	Exercises	58	
<b>3</b>	<b>Probability: multivariate models</b>	<b>61</b>	
3.1	Joint distributions for multiple random variables	61	
3.1.1	Covariance	61	
3.1.2	Correlation	62	
3.1.3	Uncorrelated does not imply independent	63	
3.1.4	Correlation does not imply causation	63	
3.1.5	Simpsons' paradox	64	
3.2	The multivariate Gaussian (normal) distribution	65	

3.2.1	Definition	65	
3.2.2	Mahalanobis distance	67	
3.2.3	Marginals and conditionals of an MVN *	68	
3.2.4	Example: Imputing missing values *	69	
3.3	Linear Gaussian systems *	69	
3.3.1	Example: inferring a latent vector from a noisy sensor	71	
3.3.2	Example: inferring a latent vector from multiple noisy sensors	72	
3.4	Mixture models	72	
3.4.1	Gaussian mixture models	73	
3.4.2	Mixtures of Bernoullis	77	
3.4.3	Gaussian scale mixtures *	77	
3.5	Probabilistic graphical models *	78	
3.5.1	Representation	79	
3.5.2	Inference	81	
3.5.3	Learning	83	
3.6	Exercises	84	
<b>4</b>	<b>Statistics</b>	<b>87</b>	
4.1	Introduction	87	
4.2	Maximum likelihood estimation (MLE)	87	
4.2.1	Definition	87	
4.2.2	Justification for MLE	88	
4.2.3	Example: MLE for the Bernoulli distribution	90	
4.2.4	Example: MLE for the categorical distribution	90	
4.2.5	Example: MLE for the univariate Gaussian	91	
4.2.6	Example: MLE for the multivariate Gaussian	92	
4.2.7	Example: MLE for linear regression	94	
4.3	Empirical risk minimization (ERM)	95	
4.3.1	Example: minimizing the misclassification rate	95	
4.3.2	Surrogate loss	96	
4.4	Other estimation methods *	96	
4.4.1	The method of moments	96	
4.4.2	Online (recursive) estimation	98	
4.5	Regularization	100	
4.5.1	Example: MAP estimation for the Bernoulli distribution	101	
4.5.2	Example: MAP estimation for the multivariate Gaussian *	102	
4.5.3	Example: weight decay	103	
4.5.4	Picking the regularizer using a validation set	105	
4.5.5	Cross-validation	105	
4.5.6	Early stopping	107	
4.5.7	Using more data	108	
4.6	Bayesian statistics *	109	
4.6.1	Conjugate priors	109	
4.6.2	The beta-binomial model	110	
4.6.3	The Dirichlet-multinomial model	117	

4.6.4	The Gaussian-Gaussian model	120
4.6.5	Beyond conjugate priors	123
4.6.6	Credible intervals	124
4.6.7	Bayesian machine learning	126
4.6.8	Computational issues	130
4.7	Frequentist statistics *	132
4.7.1	Sampling distributions	133
4.7.2	Gaussian approximation of the sampling distribution of the MLE	133
4.7.3	Bootstrap approximation of the sampling distribution of any estimator	134
4.7.4	Confidence intervals	136
4.7.5	The bias-variance tradeoff	138
4.8	Exercises	142
<b>5</b>	<b>Decision theory</b>	<b>145</b>
5.1	Bayesian decision theory	145
5.1.1	Basics	145
5.1.2	Classification problems	146
5.1.3	ROC curves	148
5.1.4	Precision-recall curves	151
5.1.5	Regression problems	153
5.1.6	Probabilistic prediction problems	154
5.2	A/B testing *	156
5.2.1	A Bayesian approach	157
5.2.2	Example	160
5.3	Bandit problems *	161
5.3.1	Contextual bandits	162
5.3.2	Markov decision processes	162
5.3.3	Exploration-exploitation tradeoff	163
5.3.4	Optimal solution	163
5.3.5	Regret	165
5.3.6	Upper confidence bounds (UCB)	166
5.3.7	Thompson sampling	168
5.3.8	Simple heuristics	168
5.4	Bayesian hypothesis testing	169
5.4.1	Example: Testing if a coin is fair	170
5.4.2	Bayesian model selection	170
5.4.3	Occam's razor	172
5.4.4	Connection between cross validation and marginal likelihood	173
5.4.5	Information criteria *	175
5.5	Frequentist decision theory	177
5.5.1	Computing the risk of an estimator	177
5.5.2	Consistent estimators	180
5.5.3	Admissible estimators	180
5.6	Empirical risk minimization	181
5.6.1	Empirical risk	181

5.6.2	Structural risk	183	
5.6.3	Cross-validation	184	
5.6.4	Statistical learning theory *	185	
5.7	Frequentist hypothesis testing *	186	
5.7.1	Likelihood ratio test	186	
5.7.2	Null hypothesis significance testing (NHST)	187	
5.7.3	p-values	188	
5.8	Exercises	190	
<b>6</b>	<b>Information theory</b>	<b>193</b>	
6.1	Entropy	193	
6.1.1	Entropy for discrete random variables	193	
6.1.2	Cross entropy	195	
6.1.3	Joint entropy	195	
6.1.4	Conditional entropy	196	
6.1.5	Perplexity	197	
6.1.6	Differential entropy for continuous random variables *	198	
6.2	Relative entropy (KL divergence) *	199	
6.2.1	Definition	199	
6.2.2	Interpretation	200	
6.2.3	Example: KL divergence between two Gaussians	200	
6.2.4	Non-negativity of KL	200	
6.2.5	KL divergence and MLE	201	
6.2.6	Forward vs reverse KL	202	
6.3	Mutual information *	203	
6.3.1	Definition	203	
6.3.2	Interpretation	203	
6.3.3	Example	204	
6.3.4	Conditional mutual information	205	
6.3.5	Normalized mutual information	205	
6.3.6	MI as a “generalized correlation coefficient”	206	
6.3.7	Data processing inequality	208	
6.3.8	Sufficient Statistics	209	
6.3.9	Fano’s inequality *	209	
6.4	Exercises	210	
<b>7</b>	<b>Linear algebra</b>	<b>213</b>	
7.1	Introduction	213	
7.1.1	Notation	213	
7.1.2	Vector spaces	216	
7.1.3	Norms of a vector and matrix	218	
7.1.4	Properties of a matrix	220	
7.1.5	Special types of matrices	222	
7.2	Matrix multiplication	226	
7.2.1	Vector-Vector Products	226	

7.2.2	Matrix-Vector Products	226	
7.2.3	Matrix-Matrix Products	227	
7.2.4	Application: manipulating data matrices	229	
7.2.5	Kronecker products *	231	
7.2.6	Einstein summation *	232	
7.3	Matrix inversion	233	
7.3.1	The inverse of a square matrix	233	
7.3.2	Schur complements *	233	
7.3.3	The matrix inversion lemma *	235	
7.3.4	Matrix determinant lemma *	235	
7.4	Eigenvalue decomposition (EVD)	236	
7.4.1	Basics	236	
7.4.2	Diagonalization	237	
7.4.3	Eigenvalues and eigenvectors of symmetric matrices	237	
7.4.4	Geometry of quadratic forms	238	
7.4.5	Standardizing and whitening data	238	
7.4.6	Power method	240	
7.4.7	Deflation	241	
7.4.8	Eigenvectors optimize quadratic forms	241	
7.5	Singular value decomposition (SVD)	241	
7.5.1	Basics	241	
7.5.2	Connection between SVD and EVD	242	
7.5.3	Pseudo inverse	243	
7.5.4	SVD and the range and null space of a matrix *	244	
7.5.5	Truncated SVD	245	
7.6	Other matrix decompositions *	246	
7.6.1	LU factorization	246	
7.6.2	QR decomposition	246	
7.6.3	Cholesky decomposition	247	
7.7	Solving systems of linear equations *	248	
7.7.1	Solving square systems	249	
7.7.2	Solving underconstrained systems (least norm estimation)	249	
7.7.3	Solving overconstrained systems (least squares estimation)	250	
7.8	Matrix calculus	251	
7.8.1	Derivatives	251	
7.8.2	Gradients	252	
7.8.3	Directional derivative	253	
7.8.4	Total derivative *	253	
7.8.5	Jacobian	253	
7.8.6	Hessian	254	
7.8.7	Gradients of commonly used functions	254	
7.9	Exercises	256	
<b>8</b>	<b>Optimization</b>	<b>257</b>	
8.1	Introduction	257	

8.1.1	Local vs global optimization	257	
8.1.2	Constrained vs unconstrained optimization	259	
8.1.3	Convex vs nonconvex optimization	259	
8.1.4	Smooth vs nonsmooth optimization	263	
8.2	First-order methods	264	
8.2.1	Descent direction	265	
8.2.2	Step size (learning rate)	266	
8.2.3	Convergence rates	268	
8.2.4	Momentum methods	269	
8.3	Second-order methods	271	
8.3.1	Newton's method	271	
8.3.2	BFGS and other quasi-Newton methods	272	
8.3.3	Trust region methods	273	
8.3.4	Natural gradient descent *	274	
8.4	Stochastic gradient descent	277	
8.4.1	Application to finite sum problems	277	
8.4.2	Example: SGD for fitting linear regression	278	
8.4.3	Choosing the step size	279	
8.4.4	Iterate averaging	279	
8.4.5	Variance reduction *	280	
8.4.6	Preconditioned SGD	281	
8.5	Constrained optimization	283	
8.5.1	Lagrange multipliers	284	
8.5.2	The KKT conditions	286	
8.5.3	Linear programming	287	
8.5.4	Quadratic programming	288	
8.5.5	Mixed integer linear programming *	289	
8.6	Proximal gradient method *	290	
8.6.1	Projected gradient descent	290	
8.6.2	Proximal operator for $\ell_1$ -norm regularizer	291	
8.6.3	Proximal operator for quantization	293	
8.7	Bound optimization *	293	
8.7.1	The general algorithm	294	
8.7.2	The EM algorithm	294	
8.7.3	Example: EM for a GMM	297	
8.7.4	Example: EM for an MVN with missing data	301	
8.8	Blackbox and derivative free optimization	304	
8.8.1	Grid search and random search	304	
8.8.2	Simulated annealing *	304	
8.8.3	Model-based blackbox optimization *	305	
8.9	Exercises	306	

<b>II</b>	<b>Linear models</b>	<b>307</b>
<b>9</b>	<b>Linear discriminant analysis</b>	<b>309</b>
9.1	Introduction	309
9.2	Gaussian discriminant analysis	309
9.2.1	Quadratic decision boundaries	310
9.2.2	Linear decision boundaries	311
9.2.3	The connection between LDA and logistic regression	311
9.2.4	Model fitting	312
9.2.5	Nearest centroid classifier	314
9.2.6	Fisher's linear discriminant analysis *	314
9.3	Naive Bayes classifiers	318
9.3.1	Example models	319
9.3.2	Model fitting	320
9.3.3	Bayesian naive Bayes	321
9.3.4	The connection between naive Bayes and logistic regression	321
9.4	Generative vs discriminative classifiers	322
9.4.1	Advantages of discriminative classifiers	322
9.4.2	Advantages of generative classifiers	323
9.4.3	Handling missing features	323
9.5	Exercises	324
<b>10</b>	<b>Logistic regression</b>	<b>325</b>
10.1	Introduction	325
10.2	Binary logistic regression	325
10.2.1	Linear classifiers	325
10.2.2	Nonlinear classifiers	326
10.2.3	Maximum likelihood estimation	328
10.2.4	Stochastic gradient descent	331
10.2.5	Perceptron algorithm	331
10.2.6	Iteratively reweighted least squares	332
10.2.7	MAP estimation	333
10.2.8	Standardization	335
10.3	Multinomial logistic regression	336
10.3.1	Linear and nonlinear classifiers	336
10.3.2	Maximum likelihood estimation	336
10.3.3	Gradient-based optimization	339
10.3.4	Bound optimization	339
10.3.5	MAP estimation	340
10.3.6	Maximum entropy classifiers	341
10.3.7	Hierarchical classification	342
10.3.8	Handling large numbers of classes	343
10.4	Preprocessing discrete input data	344
10.4.1	One-hot encoding	344
10.4.2	Feature crosses	345



10.4.3	Dealing with text	345
10.4.4	Handling missing data	348
10.5	Robust logistic regression *	349
10.5.1	Mixture model for the likelihood	349
10.5.2	Bi-tempered loss	350
10.6	Bayesian logistic regression *	352
10.6.1	Laplace approximation	352
10.6.2	Approximating the posterior predictive	354
10.7	Exercises	356
<b>11</b>	<b>Linear regression</b>	<b>359</b>
11.1	Introduction	359
11.2	Standard linear regression	359
11.2.1	Terminology	359
11.2.2	Least squares estimation	360
11.2.3	Other approaches to computing the MLE	364
11.2.4	Measuring goodness of fit	367
11.3	Ridge regression	369
11.3.1	Computing the MAP estimate	370
11.3.2	Connection between ridge regression and PCA	371
11.3.3	Choosing the strength of the regularizer	373
11.4	Robust linear regression *	373
11.4.1	Robust regression using the Student $t$ distribution	373
11.4.2	Robust regression using the Laplace distribution	375
11.4.3	Robust regression using Huber loss	376
11.4.4	Robust regression by randomly or iteratively removing outliers	377
11.5	Lasso regression	377
11.5.1	MAP estimation with a Laplace prior ( $\ell_1$ regularization)	377
11.5.2	Why does $\ell_1$ regularization yield sparse solutions?	378
11.5.3	Hard vs soft thresholding	379
11.5.4	Regularization path	381
11.5.5	Comparison of least squares, lasso, ridge and subset selection	382
11.5.6	Variable selection consistency	384
11.5.7	Group lasso	385
11.5.8	Elastic net (ridge and lasso combined)	387
11.5.9	Optimization algorithms	387
11.6	Bayesian linear regression *	389
11.6.1	Computing the posterior	389
11.6.2	Computing the posterior predictive	392
11.6.3	Empirical Bayes (Automatic relevancy determination)	394
11.7	Exercises	396
<b>12</b>	<b>Generalized linear models *</b>	<b>399</b>
12.1	Introduction	399
12.2	The exponential family	399

12.2.1	Definition	399	
12.2.2	Examples	400	
12.2.3	Log partition function is cumulant generating function		405
12.2.4	MLE for the exponential family	406	
12.2.5	Exponential dispersion family	407	
12.2.6	Maximum entropy derivation of the exponential family		407
12.3	Generalized linear models (GLMs)	408	
12.3.1	Examples	409	
12.3.2	Maximum likelihood estimation	411	
12.3.3	GLMs with non-canonical link functions		411
12.4	Probit regression	412	
12.4.1	Latent variable interpretation	412	
12.4.2	Maximum likelihood estimation	413	
12.4.3	Ordinal probit regression *	415	
12.4.4	Multinomial probit models *	415	

### III Deep neural networks 417

#### 13 Neural networks for unstructured data 419

13.1	Introduction	419	
13.2	Multilayer perceptrons (MLPs)	420	
13.2.1	The XOR problem	420	
13.2.2	Differentiable MLPs	421	
13.2.3	Activation functions	422	
13.2.4	Example models	424	
13.2.5	The importance of depth	428	
13.2.6	Connections with biology	430	
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	435	
13.3.4	Computation graphs	438	
13.4	Training neural networks	439	
13.4.1	Tuning the learning rate	440	
13.4.2	Vanishing gradient problem	441	
13.4.3	Difficulties training deep models	443	
13.4.4	Residual connections	444	
13.4.5	Batch normalization	445	
13.4.6	Parameter initialization	446	
13.5	Regularization	448	
13.5.1	Early stopping	449	
13.5.2	Weight decay	449	
13.5.3	Sparse DNNs	449	
13.5.4	Dropout	450	

13.5.5	Bayesian neural networks	451	
13.6	Other kinds of feedforward networks	451	
13.6.1	Radial basis function networks	451	
13.6.2	Mixtures of experts	453	
13.7	Exercises	456	
<b>14</b>	<b>Neural networks for images</b>	<b>459</b>	
14.1	Introduction	459	
14.2	Basics	459	
14.2.1	Convolution in 1d	459	
14.2.2	Convolution in 2d	461	
14.2.3	Convolution as matrix-vector multiplication	462	
14.2.4	Boundary conditions and strides	462	
14.2.5	Pooling layers	465	
14.2.6	Normalization layers	466	
14.2.7	Putting it altogether	467	
14.3	Image classification using CNNs	467	
14.3.1	Common datasets	467	
14.3.2	Common models	471	
14.4	Solving other discriminative vision tasks with CNNs	475	
14.4.1	Image tagging	475	
14.4.2	Object detection	476	
14.4.3	Human pose estimation	477	
14.4.4	Image segmentation	477	
14.5	Generating images by inverting CNNs *	480	
14.5.1	Converting a trained classifier into a generative model	480	
14.5.2	Image priors	480	
14.5.3	Visualizing the features learned by a CNN	482	
14.5.4	Deep Dream	483	
14.5.5	Neural style transfer	484	
14.6	Adversarial Examples *	487	
14.6.1	Whitebox (gradient-based) attacks	488	
14.6.2	Blackbox (gradient-free) attacks	489	
14.6.3	Real world adversarial attacks	490	
14.6.4	Defenses based on robust optimization	490	
14.6.5	Why models have adversarial examples	491	
<b>15</b>	<b>Neural networks for sequences</b>	<b>495</b>	
15.1	Introduction	495	
15.2	Recurrent neural networks (RNNs)	495	
15.2.1	Vec2Seq (sequence generation)	495	
15.2.2	Seq2Vec (sequence classification)	498	
15.2.3	Seq2Seq (sequence translation)	499	
15.2.4	Beam search	501	
15.2.5	Backpropagation through time	501	

15.2.6	Gating and long term memory	502
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	Seq2seq with attention	507
15.4.2	Seq2vec with attention	508
15.4.3	Attention as a soft dictionary lookup	508
15.4.4	Soft vs hard attention	510
15.5	Transformers	510
15.5.1	Self-attention	511
15.5.2	Multi-headed attention	512
15.5.3	Positional encoding	513
15.5.4	Putting it altogether	513
15.5.5	Comparing transformers, CNNs and RNNs	514
15.6	Efficient transformers *	515
15.6.1	Fixed non-learnable localized attention patterns	515
15.6.2	Learnable sparse attention patterns	516
15.6.3	Memory and recurrence methods	517
15.6.4	Low-rank and kernel methods	517

## IV Nonparametric models 521

### 16 Exemplar-based methods 523

16.1	K nearest neighbor (KNN) classification	523
16.1.1	Example	524
16.1.2	The curse of dimensionality	524
16.1.3	Reducing the speed and memory requirements	526
16.1.4	Open set recognition	526
16.2	Learning distance metrics	527
16.2.1	Linear and convex methods	528
16.2.2	Deep metric learning	529
16.2.3	Classification losses	530
16.2.4	Ranking losses	530
16.2.5	Speeding up ranking loss optimization	532
16.2.6	Other training tricks for DML	535
16.3	Kernel density estimation (KDE)	535
16.3.1	Density kernels	536
16.3.2	Parzen window density estimator	536
16.3.3	How to choose the bandwidth parameter	538
16.3.4	From KDE to KNN classification	538
16.3.5	Kernel regression	539

### 17 Kernel methods 543

17.1	Inferring functions from data	543	
17.1.1	Smoothness prior	544	
17.1.2	Inference from noise-free observations	544	
17.1.3	Inference from noisy observations	546	
17.2	Mercer kernels	546	
17.2.1	Mercer's theorem	547	
17.2.2	Some popular Mercer kernels	547	
17.3	Gaussian processes	552	
17.3.1	Noise-free observations	552	
17.3.2	Noisy observations	553	
17.3.3	Comparison to kernel regression	554	
17.3.4	Weight space vs function space	555	
17.3.5	Numerical issues	556	
17.3.6	Estimating the kernel	556	
17.3.7	GPs for classification	559	
17.3.8	Connections with deep learning	560	
17.4	Scaling GPs to large datasets	561	
17.4.1	Sparse (inducing-point) approximations	561	
17.4.2	Exploiting parallelization and kernel matrix structure	561	
17.4.3	Random feature approximation	561	
17.5	Support vector machines (SVMs)	563	
17.5.1	Large margin classifiers	563	
17.5.2	The dual problem	565	
17.5.3	Soft margin classifiers	567	
17.5.4	The kernel trick	568	
17.5.5	Converting SVM outputs into probabilities	569	
17.5.6	Connection with logistic regression	569	
17.5.7	Multi-class classification with SVMs	570	
17.5.8	How to choose the regularizer $C$	571	
17.5.9	Kernel ridge regression	572	
17.5.10	SVMs for regression	573	
17.6	Sparse vector machines	575	
17.6.1	Relevance vector machines (RVMs)	576	
17.6.2	Comparison of sparse and dense kernel methods	576	
17.7	Optimizing in function space *	578	
17.7.1	Functional analysis	578	
17.7.2	Hilbert space	580	
17.7.3	Reproducing Kernel Hilbert Space	580	
17.7.4	Representer theorem	581	
17.7.5	Kernel ridge regression revisited	583	
17.8	Exercises	583	
<b>18</b>	<b>Trees, forests, bagging and boosting</b>	<b>585</b>	
18.1	Classification and regression trees (CART)	585	
18.1.1	Model definition	585	

18.1.2	Model fitting	587	
18.1.3	Regularization	588	
18.1.4	Handling missing input features	588	
18.1.5	Pros and cons	588	
18.2	Ensemble learning	590	
18.2.1	Stacking	590	
18.2.2	Ensembling is not Bayes model averaging	591	
18.3	Bagging	591	
18.4	Random forests	592	
18.5	Boosting	593	
18.5.1	Forward stagewise additive modeling	594	
18.5.2	Quadratic loss and least squares boosting	594	
18.5.3	Exponential loss and AdaBoost	595	
18.5.4	LogitBoost	598	
18.5.5	Gradient boosting	598	
18.6	Interpreting tree ensembles	602	
18.6.1	Feature importance	602	
18.6.2	Partial dependency plots	603	

## **V Beyond supervised learning 605**

### **19 Learning with fewer labeled examples 607**

19.1	Data augmentation	607	
19.1.1	Examples	607	
19.1.2	Theoretical justification	608	
19.2	Transfer learning	608	
19.2.1	Fine-tuning	609	
19.2.2	Supervised pre-training	610	
19.2.3	Unsupervised pre-training (self-supervised learning)	611	
19.2.4	Domain adaptation	614	
19.3	Meta-learning *	614	
19.3.1	Model-agnostic meta-learning (MAML)	615	
19.4	Few-shot learning *	616	
19.4.1	Matching networks	616	
19.5	Word embeddings	618	
19.5.1	Methods based on SVD	618	
19.5.2	Word2vec	620	
19.5.3	RAND-WALK model of word embeddings	622	
19.5.4	Word analogies	623	
19.5.5	Contextual word embeddings	623	
19.6	Semi-supervised learning	628	
19.6.1	Self-training and pseudo-labeling	628	
19.6.2	Entropy minimization	630	
19.6.3	Co-training	632	

19.6.4	Label propagation on graphs	633
19.6.5	Consistency regularization	634
19.6.6	Deep generative models *	635
19.6.7	Combining self-supervised and semi-supervised learning	639
19.7	Active learning	640
19.7.1	Decision-theoretic approach	640
19.7.2	Information-theoretic approach	641
19.7.3	Batch active learning	641
19.8	Exercises	642
<b>20</b>	<b>Dimensionality reduction</b>	<b>643</b>
20.1	Principal components analysis (PCA)	643
20.1.1	Examples	643
20.1.2	Derivation of the algorithm	645
20.1.3	Computational issues	648
20.1.4	Choosing the number of latent dimensions	650
20.2	Factor analysis *	652
20.2.1	Generative model	653
20.2.2	Probabilistic PCA	654
20.2.3	EM algorithm for FA/PPCA	655
20.2.4	Unidentifiability of the parameters	657
20.2.5	Nonlinear factor analysis	659
20.2.6	Mixtures of factor analysers	660
20.2.7	Exponential family factor analysis	661
20.2.8	Factor analysis models for paired data	663
20.3	Autoencoders	666
20.3.1	Bottleneck autoencoders	666
20.3.2	Denoising autoencoders	667
20.3.3	Contractive autoencoders	668
20.3.4	Sparse autoencoders	669
20.3.5	Variational autoencoders	671
20.4	Manifold learning *	674
20.4.1	What are manifolds?	675
20.4.2	The manifold hypothesis	676
20.4.3	Approaches to manifold learning	676
20.4.4	Multi-dimensional scaling (MDS)	677
20.4.5	Isomap	680
20.4.6	Kernel PCA	681
20.4.7	Maximum variance unfolding (MVU)	683
20.4.8	Local linear embedding (LLE)	683
20.4.9	Laplacian eigenmaps	685
20.4.10	t-SNE	687
20.5	Exercises	691
<b>21</b>	<b>Clustering</b>	<b>695</b>

21.1	Introduction	695	
21.1.1	Evaluating the output of clustering methods	695	
21.2	Hierarchical agglomerative clustering	697	
21.2.1	The algorithm	698	
21.2.2	Example	700	
21.3	K means clustering	701	
21.3.1	The algorithm	702	
21.3.2	Examples	702	
21.3.3	Vector quantization	703	
21.3.4	The K-means++ algorithm	705	
21.3.5	The K-medoids algorithm	705	
21.3.6	Speedup tricks	706	
21.3.7	Choosing the number of clusters $K$	706	
21.4	Clustering using mixture models	710	
21.4.1	Mixtures of Gaussians	710	
21.4.2	Mixtures of Bernoullis	714	
21.5	Spectral clustering *	715	
21.5.1	Normalized cuts	715	
21.5.2	Eigenvectors of the graph Laplacian encode the clustering	715	
21.5.3	Example	716	
21.5.4	Connection with other methods	717	
21.6	Biclustering *	717	
21.6.1	Basic biclustering	718	
21.6.2	Nested partition models (Crosscat)	718	
<b>22</b>	<b>Recommender systems</b>	<b>721</b>	
22.1	Explicit feedback	721	
22.1.1	Datasets	721	
22.1.2	Collaborative filtering	722	
22.1.3	Matrix factorization	723	
22.1.4	Autoencoders	725	
22.2	Implicit feedback	726	
22.2.1	Bayesian personalized ranking	727	
22.2.2	Factorization machines	727	
22.2.3	Neural matrix factorization	728	
22.3	Leveraging side information	729	
22.4	Exploration-exploitation tradeoff	730	
<b>23</b>	<b>Graph embeddings *</b>	<b>733</b>	
23.1	Introduction	733	
23.2	Graph Embedding as an Encoder/Decoder Problem	734	
23.3	Shallow graph embeddings	736	
23.3.1	Unsupervised embeddings	737	
23.3.2	Distance-based: Euclidean methods	737	
23.3.3	Distance-based: non-Euclidean methods	738	



23.3.4	Outer product-based: Matrix factorization methods	738
23.3.5	Outer product-based: Skip-gram methods	739
23.3.6	Supervised embeddings	740
23.4	Graph Neural Networks	741
23.4.1	Message passing GNNs	741
23.4.2	Spectral Graph Convolutions	743
23.4.3	Spatial Graph Convolutions	743
23.4.4	Non-Euclidean Graph Convolutions	745
23.5	Deep graph embeddings	745
23.5.1	Unsupervised embeddings	745
23.5.2	Semi-supervised embeddings	748
23.6	Applications	749
23.6.1	Unsupervised applications	749
23.6.2	Supervised applications	751

**Appendices 753**

**VI Appendix 755**

**A Notation 757**

A.1	Introduction	757
A.2	Common mathematical symbols	757
A.3	Functions	758
A.3.1	Common functions of one argument	758
A.3.2	Common functions of two arguments	758
A.3.3	Common functions of > 2 arguments	758
A.4	Linear algebra	759
A.4.1	General notation	759
A.4.2	Vectors	759
A.4.3	Matrices	759
A.4.4	Matrix calculus	760
A.5	Optimization	760
A.6	Probability	761
A.7	Information theory	761
A.8	Statistics and machine learning	761
A.8.1	Supervised learning	762
A.8.2	Unsupervised learning and generative models	762
A.8.3	Bayesian inference	762
A.9	Abbreviations	763