

概要目次

I 編	前提知識	1
1 章	導 入	3
2 章	数学的基礎	35
3 章	言語学の要点	73
4 章	コーパスに基づく研究	105
II 編	語	135
5 章	連 語	137
6 章	統計的推論：スパースなデータ上の n -グラムモデル	171
7 章	語義の曖昧性解消	203
8 章	語彙獲得	233
III 編	文 法	277
9 章	マルコフモデル	279
10 章	品詞のタグ付け	301
11 章	確率文脈自由文法	337
12 章	確率的構文解析	359
IV 編	応用と技法	407
13 章	統計的アライメントと機械翻訳	409
14 章	クラスタリング	439
15 章	情報検索におけるいくつかの話題	471
16 章	テキスト分類	513

目次

表目次	xiii
図目次	xviii
記法一覧	xxii
まえがき	xxv
本書の使い方	xxix
I 編 前提知識	1
1 章 導 入	3
1.1 言語に対する合理主義的方法論と経験主義的方法論	5
1.2 科学的意義	7
1.2.1 言語学が答えるべき質問	8
1.2.2 カテゴリカルでない言語現象	11
1.2.3 確率的現象としての言語と認知	14
1.3 言語の曖昧性：なぜ自然言語処理は困難なのか	16
1.4 汚れ仕事	18
1.4.1 言語資源	18
1.4.2 単語数	19
1.4.3 ジップの法則など	22

1.4.4	連語	27
1.4.5	コンコーダンス	30
1.5	さらに学ぶために	31
1.6	練習問題	33

2章 数学的基礎

35

2.1	確率論の基礎	36
2.1.1	確率空間	36
2.1.2	条件付き確率と独立性	38
2.1.3	ベイズの定理	39
2.1.4	確率変数	41
2.1.5	期待値と分散	42
2.1.6	記法	43
2.1.7	結合分布と条件付き分布	43
2.1.8	P の決定	44
2.1.9	標準的な分布	45
2.1.10	ベイズ統計	49
2.1.11	練習問題	53
2.2	情報理論の要点	54
2.2.1	エントロピー	54
2.2.2	結合エントロピーと条件付きエントロピー	57
2.2.3	相互情報量	60
2.2.4	雑音のある通信路モデル	61
2.2.5	相対エントロピーとカルバック・ライブラー・ダイバージェンス	64
2.2.6	言語との関係：交差エントロピー	65
2.2.7	英語のエントロピー	69
2.2.8	パープレキシティ	70
2.2.9	練習問題	71
2.3	さらに学ぶために	71

3章 言語学の要点

73

3.1	品詞と形態論	73
3.1.1	名詞と代名詞	75
3.1.2	名詞に伴う語：限定詞と形容詞	78

3.1.3	動詞	80	
3.1.4	その他の品詞	82	
3.2	句構造	84	
3.2.1	句構造文法	86	
3.2.2	依存：項と付加語	91	
3.2.3	X' 理論	96	
3.2.4	句構造の曖昧性	97	
3.3	意味論と語用論	99	
3.4	その他の分野	102	
3.5	さらに学ぶために	103	
3.6	練習問題	103	
4 章	コーパスに基づく研究		105
4.1	準備	106	
4.1.1	計算機	106	
4.1.2	コーパス	106	
4.1.3	ソフトウェア	108	
4.2	テキストの観察	110	
4.2.1	低いレベルの書式に関する問題	111	
4.2.2	トークン化：語とは何か	112	
4.2.3	形態論	119	
4.2.4	文	121	
4.3	データのマークアップ	123	
4.3.1	マークアップの枠組み	123	
4.3.2	文法的タグ付け	125	
4.4	さらに学ぶために	131	
4.5	練習問題	133	
II 編	語		135
5 章	連語		137
5.1	頻度	139	
5.2	平均と分散	142	
5.3	仮説検定	147	
5.3.1	t 検定	147	

5.3.2	差の仮説検定	150
5.3.3	Pearson のカイ二乗検定	152
5.3.4	尤度比	155
5.4	相互情報量	159
5.5	連語とは何か	164
5.6	さらに学ぶために	167

6 章 統計的推論：スパースなデータ上の n -グラムモデル 171

6.1	ビン：同値類の形成	171
6.1.1	信頼性と弁別性	171
6.1.2	n -グラムモデル	172
6.1.3	n -グラムモデルの構築	174
6.2	統計的推定	175
6.2.1	最尤推定 (Most Likelihood Estimator: MLE)	176
6.2.2	Laplace の法則, Lidstone の法則, Jeffreys-Perks の法則	180
6.2.3	ヘルドアウト推定	183
6.2.4	交差検証 (削除推定)	188
6.2.5	Good-Turing の推定	189
6.2.6	短い補足	192
6.3	推定値を組み合わせる	194
6.3.1	単純な線形補間	194
6.3.2	Katz のバックオフ	195
6.3.3	一般的な線形補間	196
6.3.4	短い注意書き	198
6.3.5	Austen のための言語モデル	199
6.4	結論	200
6.5	さらに学ぶために	200
6.6	練習問題	201

7 章 語義の曖昧性解消 203

7.1	方法論に関する準備	206
7.1.1	教師あり学習と教師なし学習	206
7.1.2	擬似語	206
7.1.3	性能の上限と下限	207

7.2	教師あり曖昧性解消	208
7.2.1	ベイズ分類	209
7.2.2	情報理論的アプローチ	212
7.3	辞書に基づく曖昧性解消	214
7.3.1	意味の定義に基づく曖昧性解消	214
7.3.2	シソーラスに基づく曖昧性解消	216
7.3.3	第二言語のコーパスにおける翻訳に基づく曖昧性解消	219
7.3.4	談話内で意味は一つ，同一の連語で意味は一つ	220
7.4	教師なし曖昧性解消	223
7.5	単語の意味とは何か?	226
7.6	さらに学ぶために	230
7.7	練習問題	231
8 章 語彙獲得		233
8.1	評価指標	235
8.2	動詞の下位範疇化	238
8.3	付加の曖昧性	245
8.3.1	Hindle and Rooth(1993)の方法	247
8.3.2	前置詞の付加先決定に関する総合的なコメント	250
8.4	選択選好	253
8.5	意味的類似性	259
8.5.1	ベクトル空間尺度	261
8.5.2	確率的尺度	266
8.6	統計的自然言語処理における語彙獲得の役割	271
8.7	さらに学ぶために	274
III 編 文 法		277
9 章 マルコフモデル		279
9.1	マルコフモデル	280
9.2	隠れマルコフモデル	282
9.2.1	なぜ HMM を用いるのか	283
9.2.2	HMM の一般形	285
9.3	HMM についての三つの基本的な問題	286

9.3.1	観測の確率を求める	287
9.3.2	最適な状態系列を求める	291
9.3.3	三つ目の問題：パラメータの推定	293
9.4	HMM：実装，性質，変種	296
9.4.1	実装	296
9.4.2	HMM における変種	297
9.4.3	複数の観測入力	298
9.4.4	パラメータの値の初期化	298
9.5	さらに学ぶために	299

10 章 品詞のタグ付け

301

10.1	タグ付けのための情報源	303
10.2	マルコフモデルによるタグ付け器	304
10.2.1	確率モデル	304
10.2.2	ビタビアルゴリズム	308
10.2.3	さまざまなバリエーション	310
10.3	隠れマルコフモデルによるタグ付け器	315
10.3.1	HMM を品詞タグ付けに適用する	315
10.3.2	HMM 学習における初期化の影響	317
10.4	変換に基づくタグの学習	319
10.4.1	変換規則	320
10.4.2	学習アルゴリズム	321
10.4.3	ほかのモデルとの関係	322
10.4.4	オートマトン	324
10.4.5	変換に基づくタグ付けに関するまとめ	326
10.5	別の方法，英語以外の言語	327
10.5.1	タグ付けに対する別のアプローチ	327
10.5.2	英語以外の言語	328
10.6	タグ付けの正解率とタグ付け器の適用先	328
10.6.1	タグ付けの正解率	328
10.6.2	タグ付けの適用先	331
10.7	さらに学ぶために	334
10.8	練習問題	336

11 章 確率文脈自由文法	337
11.1 PCFG のいくつかの特徴	341
11.2 PCFG の三つの基本的な問題	343
11.3 系列の確率	347
11.3.1 内側確率により計算する	347
11.3.2 外側確率により計算する	348
11.3.3 最も尤もらしい文の構文木を求める	350
11.3.4 PCFG の訓練	352
11.4 内側外側アルゴリズムの問題点	355
11.5 さらに学ぶために	356
11.6 練習問題	357
12 章 確率的構文解析	359
12.1 いくつかの概念	360
12.1.1 曖昧性解消のための構文解析	360
12.1.2 ツリーバンク	363
12.1.3 構文解析モデル vs. 言語モデル	365
12.1.4 PCFG の独立性の仮定を弱める	367
12.1.5 構文木の確率と導出確率	371
12.1.6 方法は一つだけではない	373
12.1.7 句構造文法と依存文法	377
12.1.8 評価	380
12.1.9 等価なモデル	385
12.1.10 構文解析器を構築する：探索手段	387
12.1.11 幾何平均の利用	390
12.2 いくつかのアプローチ	391
12.2.1 語彙化されていないツリーバンク文法	391
12.2.2 導出履歴を用いる語彙化されたモデル	395
12.2.3 依存構造に基づくモデル	398
12.2.4 議論	401
12.3 さらに学ぶために	402
12.4 練習問題	404

目次	xi
IV 編 応用と技法	407
13 章 統計的アライメントと機械翻訳	409
13.1 テキストアライメント 412	
13.1.1 文や段落のアライメント 413	
13.1.2 長さに基づく方法 416	
13.1.3 信号処理技術に基づくオフセットのアライメント 420	
13.1.4 語彙的文アライメント手法 423	
13.1.5 まとめ 428	
13.1.6 練習問題 428	
13.2 語のアライメント 429	
13.3 統計的機械翻訳 430	
13.4 さらに学ぶために 436	
14 章 クラスタリング	439
14.1 階層的クラスタリング 445	
14.1.1 単一リンククラスタリングと完全リンククラスタリング 446	
14.1.2 群平均凝集型クラスタリング 450	
14.1.3 応用：言語モデルの改善 452	
14.1.4 トップダウンクラスタリング 455	
14.2 非階層的クラスタリング 456	
14.2.1 K 平均法 458	
14.2.2 EM アルゴリズム 461	
14.3 さらに学ぶために 468	
14.4 練習問題 469	
15 章 情報検索におけるいくつかの話	471
15.1 情報検索に関する背景知識 472	
15.1.1 情報検索システムに共通する設計の特徴点 474	
15.1.2 評価尺度 476	
15.1.3 確率的順位付け原理 479	
15.2 ベクトル空間モデル 480	
15.2.1 ベクトルの類似度 481	

15.2.2	タームの重み付け	482	
15.3	タームの分布のモデル	484	
15.3.1	ポアソン分布	485	
15.3.2	2-ポアソンモデル	488	
15.3.3	K 混合分布	489	
15.3.4	逆文書頻度	491	
15.3.5	残差逆文書頻度	492	
15.3.6	ターム分布モデルの利用	493	
15.4	潜在意味インデキシング	493	
15.4.1	最小二乗法	495	
15.4.2	特異値分解	497	
15.4.3	情報検索における潜在意味インデキシング	501	
15.5	談話分割	504	
15.5.1	テキストタイリング	505	
15.6	さらに学ぶために	508	
15.7	練習問題	510	
16 章	テキスト分類		513
16.1	決定木	515	
16.2	最大エントロピーモデル	525	
16.2.1	一般化反復スケーリング法	527	
16.2.2	テキスト分類への応用	530	
16.3	パーセプトロン	532	
16.4	k 最近傍分類	538	
16.5	さらに学ぶために	540	
	簡易統計表		543
	参考文献		545
	訳者あとがき		583
	索引		587

表 目 次

表 1.1	トム・ソーヤにおいて多く現れている語.	20
表 1.2	トム・ソーヤにおける語タイプの出現頻度の頻度.	21
表 1.3	トム・ソーヤを用いたジップの法則の実証的評価.	23
表 1.4	ニューヨークタイムズで最も頻繁に現れるバイグラム連語.	28
表 1.5	フィルタリング後の頻出バイグラム.	29
表 2.1	二つの理論の尤度比.	52
表 2.2	復号化問題としての統計的自然言語処理.	64
表 3.1	名詞の一般的な屈折.	76
表 3.2	英語における代名詞の語形.	77
表 3.3	一般に動詞に記される特徴.	81
表 4.1	電子コーパスを提供する主な組織とその連絡先 URL.	107
表 4.2	<i>The Economist</i> の一つの号に現れた電話番号のさまざまな形式.	118
表 4.3	ニュース記事テキストの文の長さ.	123
表 4.4	いくつかのタグセットの大きさ.	126
表 4.5	異なるタグセットの比較：形容詞，副詞，接続詞，限定詞，名詞， 代名詞関連のタグ.	127
表 4.6	異なるタグセットの比較：動詞，前置詞，句読点，記号関連のタグ.	128
表 5.1	連語を見つける：頻度 $C(\cdot)$ はコーパス中の頻度を表す.	140
表 5.2	連語フィルターのための品詞（タグ）のパターン.	140
表 5.3	連語を見つける：Justeson と Katz の品詞フィルター.	141
表 5.4	‘strong w ’ と ‘powerful w ’ のパターンに最も多く出現する名詞 w .	141

表 5.5	平均と分散に基づいて連語を見つける.	146
表 5.6	連語を見つける : 出現頻度 20 のバイグラム 10 個に適用した t 検定.	150
表 5.7	有意に高頻度で <i>powerful</i> とともに出現する語 (上側の 10 語), および <i>strong</i> とともに出現する語 (下側の 10 語).	151
表 5.8	<i>new</i> と <i>companies</i> の出現の依存性を表す 2×2 の表.	152
表 5.9	対訳コーパスにおける <i>vache</i> と <i>cow</i> の対応関係.	154
表 5.10	χ^2 を用いた相異なるコーパスにおける単語の独立性の検定.	154
表 5.11	Dunning の尤度比検定を計算する方法.	156
表 5.12	<i>powerful</i> のバイグラムのうち, Dunning の尤度比テストで最高スコアとなったもの.	157
表 5.13	Damerau の頻度比検定.	158
表 5.14	連語を見つける : 出現頻度頻度が 20 である 10 個のバイグラムを相互情報量の大きい順に並べたもの.	160
表 5.15	対訳形式になった Hansard コーパスにおける <i>chambre</i> と <i>house</i> , および <i>communes</i> と <i>house</i> の対応.	161
表 5.16	データスパースネスに起因する相互情報量の問題.	162
表 5.17	Cover and Thomas (1991) および Fano (1961) におけるさまざまな相互情報量の定義.	163
表 5.18	BBI Combinatory Dictionary of English における <i>strength</i> , および <i>power</i> に関する連語.	166
表 6.1	n -グラムモデルに対するパラメータ数の増加.	173
表 6.2	統計的推定の章における記法.	176
表 6.3	<i>Persuasion</i> の一つの文の直後に出現する場合の各単語の確率.	179
表 6.4	Church and Gale (1991a) にある AP データの推定頻度.	181
表 6.5	<i>was</i> に後続する単語の期待尤度推定.	183
表 6.6	二つのシステムの性能を比較するための t 検定の利用.	187
表 6.7	Austen コーパスにおけるバイグラムとトライグラムの「頻度の頻度」の分布の一部.	191
表 6.8	バイグラムに対する Good-Turing の推定値 : 調整された頻度と確率.	192
表 6.9	<i>Persuasion</i> における Good-Turing によるバイグラム頻度の推定.	192
表 6.10	<i>Persuasion</i> で評価した Good-Turing 推定によるバックオフ言語モデル.	200
表 6.11	さまざまな言語モデルに応じたテスト用の節の確率推定.	200

表目次	xv
表 7.1 本章で使用する記号の意味.	208
表 7.2 ベイズ分類で用いる <i>drug</i> の二つの語義に対する手がかり.	211
表 7.3 三つの曖昧なフランス語の単語の語義推定に対して有用な手がかり.	212
表 7.4 <i>ash</i> の二つの語義.	215
表 7.5 Lesk のアルゴリズムによる <i>ash</i> の多義解消.	215
表 7.6 シソーラスに基づく曖昧性解消のいくつかの結果.	218
表 7.7 第二言語コーパスを用いて <i>interest</i> をどのように曖昧性を解消するか.	219
表 7.8 談話内単一意味制約の例.	221
表 7.9 教師なし曖昧性解消の結果.	226
表 8.1 F 値と正解率は異なった目的関数である.	238
表 8.2 いくつかの下位範疇化フレームと動詞, および文の例.	239
表 8.3 Manning のシステムにより学習された下位範疇化フレーム.	243
表 8.4 前置詞句の付加の曖昧性を解消する単純なモデルが失敗する例.	247
表 8.5 選択選好の強度 (SPS).	256
表 8.6 結合の強さは動詞のとる目的語として尤もらしいものとそうでないものを区別する.	257
表 8.7 二値ベクトルに対する類似性尺度.	263
表 8.8 意味的類似性の尺度としてのコサイン値.	266
表 8.9 確率分布の間の (非-) 類似性の尺度.	267
表 8.10 LOB コーパスに出現する単語で OALD 辞書にカバーされていない単語の種別.	272
表 9.1 HMM の表記.	286
表 9.2 $O = \{\text{lem, ice_t, cola}\}$ に対する変数の計算.	291
表 10.1 英語のタグ付けでよく用いられる品詞.	302
表 10.2 タグ付けにおける記法上の規約.	305
表 10.3 ブラウンコーパス中のタグの理想化された遷移回数.	307
表 10.4 ブラウンコーパスにおけるいくつかの単語に対する理想化されたタグの生起回数.	307
表 10.5 タグ付けにおける未知語に対する確率割り当ての例.	311
表 10.6 HMM におけるパラメータの初期化手段.	317
表 10.7 Brill の変換に基づくタグ付け器におけるトリガ環境.	320
表 10.8 変換に基づくタグ付けにおいて学習される変換規則の例.	321
表 10.9 確率的なタグ付けで頻出する誤りの例.	330

表 10.10	品詞タグ付けにおける混同行列の一部.	331
表 11.1	本章における PCFG に関する記法.	339
表 11.2	シンプルな確率文脈自由文法 (PCFG).	340
表 11.3	内側確率の計算.	348
表 12.1	ペンツリーバンクにおける句カテゴリの略記.	365
表 12.2	いくつかの動詞においてよく現れる下位範疇化フレーム (VP を展開した部分木) の頻度.	368
表 12.3	よく見られるいくつかの展開規則による NP が主語もしくはは目的語として現れる割合.	370
表 12.4	よく見られるいくつかの展開規則による NP が VP 内部において第一目的語もしくは第二目的語として現れる割合.	370
表 12.5	異なった句構造のスタイルにおける前置詞付加の誤りに対する精度, 再現率の結果.	384
表 12.6	統計的構文解析システムの比較.	401
表 13.1	文アライメントに関する論文.	416
表 14.1	異なるクラスタリングアルゴリズムの特徴のまとめ.	444
表 14.2	クラスタリングの章で用いられる記号.	444
表 14.3	クラスタリングで用いられる類似度関数.	446
表 14.4	K 平均クラスタリングの例.	460
表 14.5	混合ガウス分布の一例.	463
表 15.1	英語の小規模なストップリスト.	475
表 15.2	順位付けの評価の例.	476
表 15.3	情報検索においてタームの重み付けに一般的に用いられる三つの量.	482
表 15.4	例となるコーパスにおける二つの語のコレクション頻度と文書頻度.	483
表 15.5	tf.idf による重み付け方式の構成要素.	484
表 15.6	ニューヨークタイムズコーパスにおける 6 単語に対する文書頻度 (df) とコレクション頻度 (cf).	487
表 15.7	六つの語について, k 回出現している文書の実際の数と推定した数.	490
表 15.8	内容の類似度計算における共起の利用の例.	493
表 15.9	文書の相関行列 $E^T E$.	500

表目次	xvii
表 16.1 自然言語処理における分類タスクの事例.	513
表 16.2 二値分類器を評価するための分割表.	515
表 16.3 図 16.3 に示された文書 11 の表現形.	518
表 16.4 分割基準としての情報利得の例.	520
表 16.5 ロイターのカテゴリ “earnings” (決算) に対する決定木の分割表.	522
表 16.6 式 (16.4) の形式における最大エントロピー分布の一例.	529
表 16.7 経験分布の一つで, 対応する最大エントロピー分布が表 16.6 のものであるもの.	529
表 16.8 ロイターのカテゴリ “earnings” (決算) に対する最大エントロピーモデルにおける素性の重み.	530
表 16.9 テストセットにおける表 16.8 に対応する分布に対する分類結果.	531
表 16.10 “earnings” カテゴリに対するパーセプトロン.	535
表 16.11 表 16.10 のパーセプトロンに対するテストセットにおける分類結果.	536
表 16.12 “earnings” カテゴリに対する 1 最近傍法に基づく分類器の分類結果.	539

目 次

図 1.1	ジップの法則.	24
図 1.2	マンデルブロの式.	25
図 1.3	語 <i>showed</i> の Key Word In Context (KWIC) 表示.	30
図 1.4	トム・ソーヤにおける <i>showed</i> の統語フレーム.	31
図 2.1	条件付き確率 $P(A B)$ の計算を描いた図式.	38
図 2.2	二つのサイコロの目の和についての確率変数 X .	41
図 2.3	二項分布の二つの例: $b(r; 10, 0.7)$ と $b(r; 10, 0.1)$.	47
図 2.4	正規分布曲線の例: $n(x; 0, 1)$ と $n(x; 1.5, 2)$.	48
図 2.5	偏りのあるコインのエントロピー.	57
図 2.6	相互情報量 I とエントロピー H との関係.	60
図 2.7	雑音のある通信路モデル.	62
図 2.8	対称的な二値通信路.	62
図 2.9	言語学における雑音のある通信路モデル.	63
図 3.1	句構造の再帰的な展開の例.	90
図 3.2	前置詞句付加の曖昧性の例.	98
図 4.1	経験的な文境界検出アルゴリズム.	122
図 4.2	いくつかの異なるタグセットでタグ付けされた文.	126
図 5.1	二つの単語が少し離れたバイグラムを捉えるための 3 単語の連語窓の利用.	143
図 5.2	<i>strong</i> から相対的に 3 語までの位置に関するヒストグラム.	145

目 次	xix
図 7.1 ベイズ曖昧性解消.	211
図 7.2 曖昧性解消のための特徴を選ぶフリップフロップアルゴリズム.	212
図 7.3 Lesk の辞書に基づく曖昧性解消.	215
図 7.4 シソーラスに基づく曖昧性解消.	216
図 7.5 シソーラスに基づく適応的な曖昧性解消.	217
図 7.6 第二言語のコーパスに基づく曖昧性解消.	220
図 7.7 「同一連語単一意味制約」, および「談話内単一意味制約」に基づく曖昧性解消.	222
図 7.8 単語の意味クラスタリングに対する EM アルゴリズム.	225
図 8.1 精度と再現率の尺度を考えるための図.	236
図 8.2 複雑な文における付加先.	251
図 8.3 文書-単語行列 A .	262
図 8.4 文書-単語行列 B .	262
図 8.5 修飾語-被修飾語行列 C .	262
図 9.1 マルコフモデルの一例.	281
図 9.2 無茶な飲料販売機. その状態と状態遷移確率.	283
図 9.3 線形補間された言語モデルのための HMM の一部.	285
図 9.4 マルコフ過程のプログラム.	286
図 9.5 トレリスアルゴリズム.	288
図 9.6 トレリスアルゴリズム: 一つのノードにおける前向き確率の計算をクローズアップしたもの.	289
図 9.7 アークをたどる確率.	294
図 10.1 可視的マルコフモデルによるタグ付け器を訓練するアルゴリズム.	307
図 10.2 可視的マルコフモデルによるタグ付け器のタグ付けアルゴリズム.	309
図 10.3 変換に基づくタグ付けの学習アルゴリズム.	322
図 11.1 二つの構文木とその確率, およびその和としての文の確率.	340
図 11.2 確率的正規文法 (PRG).	345
図 11.3 PCFG における内側確率 β , 外側確率 α .	346
図 12.1 単純化された単語ラティス.	360
図 12.2 ペンツリーバンクにおける構文木.	364
図 12.3 同じ構文木に対する二つの CFG 導出.	371
図 12.4 左隅スタック構文解析器.	374
図 12.5 部分木を依存関係に分解する.	379

図 12.6	PARSEVAL 指標の例.	382
図 12.7	括弧交差数の概念.	383
図 12.8	ペンツリーとほかの木構造の比較.	384
図 13.1	機械翻訳に対する異なる戦略.	410
図 13.2	アライメント (位置合わせ) と対応付け.	414
図 13.3	アライメントのコストの計算.	418
図 13.4	ドットプロットの例.	421
図 13.5	探索の範囲となる枕形の包絡.	425
図 13.6	機械翻訳における雑音のある通信路モデル.	430
図 14.1	22 個の英語の頻出単語に対する単一リンク法でのクラスタリング 結果を樹形図 (デンドログラム) で示したもの.	440
図 14.2	ボトムアップ型階層的クラスタリング (Bottom-up hierarchical clustering).	445
図 14.3	トップダウン型階層的クラスタリング (Top-down hierarchical clustering).	446
図 14.4	平面上の点の集まり.	447
図 14.5	図 14.4 における点群に対する中間的なクラスタリング.	447
図 14.6	図 14.4 における点群に対する単一リンククラスタリング.	447
図 14.7	図 14.4 における点群に対する完全リンククラスタリング.	449
図 14.8	K 平均クラスタリングアルゴリズム.	458
図 14.9	K 平均アルゴリズムにおける 1 回の反復.	459
図 14.10	ソフトクラスタリングに EM アルゴリズムを用いる例.	461
図 15.1	あるインターネット検索エンジンにおける ‘“glass pyramid” Pei Louvre’ の検索結果.	473
図 15.2	精度-再現率曲線の二つの事例.	478
図 15.3	2 次元のベクトル空間の例.	480
図 15.4	ポアソン分布.	486
図 15.5	ターム-文書行列 A の例.	494
図 15.6	次元圧縮.	494
図 15.7	線形回帰の例.	496
図 15.8	図 15.5 の行列に SVD を施した結果の行列 T .	498
図 15.9	図 15.5 の行列に SVD を施した結果の特異値行列.	498
図 15.10	図 15.5 の行列に SVD を施した結果の行列 D^T .	498
図 15.11	特異値による大きさの調整, ならびに 2 次元への圧縮を行った 後の文書行列 $B_{2 \times d} = S_{2 \times 2} D^T_{2 \times d}$.	500

目 次	xxi
図 15.12 話題境界同定における結束性スコアに関する三つの分布.	506
図 16.1 決定木の一例.	516
図 16.2 図 16.1 の木の一部に対する幾何学的な解釈.	516
図 16.3 ロイターのニュース記事における話題カテゴリ “earnings” (決算) の例.	517
図 16.4 決定木の枝刈り.	521
図 16.5 分類正解率は利用可能な訓練データの量に依存する.	523
図 16.6 音韻規則学習の領域におけるデータを決定木が如何に非効率的に 利用するのかという事例.	524
図 16.7 パーセプトロン学習アルゴリズム.	533
図 16.8 パーセプトロンの学習アルゴリズムにおける, 誤り訂正の一ステップ.	534
図 16.9 パーセプトロンの幾何学的な解釈.	536

記法一覧

\cup	集合の和, 結び (union)
\cap	集合の積, 交わり (intersection)
$A - B, A \setminus B$	集合の差 (difference)
\bar{A}	集合 A の補集合 (complement)
\emptyset	空集合 (empty set)
$2^A, \mathcal{P}(A)$	集合 A のべき集合 (power set)
$ A $	集合 A の要素数, デカルト数 (cardinality)
\sum	和 (sum)
\prod	積 (product)
$p \Rightarrow q$	p が q を含意 (imply) する, 論理的推論 (logical inference)
$p \Leftrightarrow q$	p と q が 論理的同値 (logically equivalent) である
$\stackrel{\text{def}}{=}$	であると定義する (“=” が曖昧なときにのみ用いる)
\mathbb{R}	実数 (real numbers) の集合
\mathbb{N}	自然数 (natural numbers) の集合
$n!$	n の階乗 (factorial)
∞	無限 (infinity)
$ x $	数 x の絶対値 (absolute value)
\ll	より, 非常に小さい
\gg	より, 非常に大きい
$f: A \rightarrow B$	A に属する値から B への関数 f
$\max f$	f の最大値 (maximum value)

$\min f$	f の最小値 (minimum value)
$\arg \max f$	f がその最大値をとるような引数
$\arg \min f$	f がその最小値をとるような引数
$\lim_{x \rightarrow \infty} f(x)$	x を無限に近づけた際の f の極限 (limit)
$f \propto g$	f が g に比例 (proportional) する
∂	偏微分 (partial derivative)
\int	積分 (integral)
$\log a$	a の対数 (logarithm)
$\exp(x), e^x$	指数関数 (exponential function)
$\lceil a \rceil$	$i \geq a$ である中で最も小さい自然数 i
\vec{x}	実数からなるベクトル: $\vec{x} \in \mathbb{R}^n$
$ \vec{x} $	\vec{x} のユークリッド長 (Euclidean length)
$\vec{x} \cdot \vec{y}$	\vec{x} と \vec{y} との内積 (dot product)
$\cos(\vec{x}, \vec{y})$	\vec{x} と \vec{y} とがなす角度のコサイン値 (cosine)
c_{ij}	行列 (matrix) C の i 行 (row) j 列 (column) の要素 (element)
C^T	行列 C の転置行列 (transpose)
\hat{X}	X の推定値 (estimate)
$E(X)$	X の期待値 (expectation)
$\text{Var}(X)$	X の分散 (variance)
μ	平均 (mean)
σ	標準偏差 (standard deviation)
\bar{x}	標本の平均 (sample mean)
s^2	標本の分散 (sample variance)
$P(A B)$	B で条件付けられた (conditional on) A の確率 (probability)
$X \sim p(x)$	p に従って分布する確率変数 (random variable) X
$b(r; n, p)$	二項分布 (binomial distribution)
$\binom{n}{r}$	組合せ (combination) もしくは二項係数 (binomial coefficient) (n 個から r 個を選ぶ, 選び方の数)
$n(x; \mu, \sigma)$	正規分布 (normal distribution)
$H(X)$	エントロピー (entropy)

$I(X; Y)$	相互情報量 (mutual information)
$D(p \parallel q)$	カルバック・ライブラー・ダイバージェンス (Kullback-Leibler (KL) divergence)
$C(\cdot)$	括弧内の事物の数
f_u	u の相対頻度 (relative frequency)
$w_{ij}, w_{(i)(j)}$	語 (words) w_i, w_{i+1}, \dots, w_j
$w_{i,j}$	w_{ij} に同じ
w_i, \dots, w_j	w_{ij} に同じ
$O(n)$	アルゴリズムの時間的複雑さ
*	非文法的 (ungrammatical) な文 (sentence) もしくは句 (phrase), あるいは 不適格 (ill-formed) な語 (word)
?	文法的 (grammatical) であるかの境界線上にある文, あるいは 受け入れられる (acceptable) かの境界線上にある句
<i>iff</i> が成り立つとき, かつそのときに限り

注：以下に示すように、いくつかの章はそこで用いられる記号についての独立した記法一覧の表を有している。表 6.2 (統計的推論), 表 7.1 (語義曖昧性解消), 表 9.1 (マルコフ・デル), 表 10.2 (タグ付け) 表 11.1 (確率文脈自由文法), 表 14.2 (クラスタリング)。