

目 次

第1章	データ分析のプロセス	1
1.1	データ分析で直面する課題	1
1.1.1	ビジネス目標実現に必要なデータ分析の定義	1
1.1.2	データ加工の煩雑さ	1
1.1.3	欠損値や外れ値への対処	2
1.1.4	不均衡データの扱い	4
1.2	データ分析のフレームワーク	5
1.3	CRISP-DM	5
1.4	KDD プロセス	7
1.5	本書におけるデータやパッケージの利用方針	8
第2章	基本的なデータ操作	10
2.1	データの入出力	10
2.1.1	テキストファイルの入出力	10
2.1.2	Excel形式のデータの入出力	13
2.1.3	リレーショナルデータベースとの入出力	14
2.2	データフレームのハンドリング	17
2.2.1	データの加工・集計	17
2.2.2	テーブルの結合	25
2.2.3	テーブルの形式の変換	26
2.3	データテーブルのハンドリング	29
2.3.1	データの読み込み	29
2.3.2	要素の抽出	30
第3章	前処理・変換	34
3.1	データの記述・要約	35
3.1.1	要約統計量の算出	35
3.1.2	データ項目間の関係の理解	36

viii 目次

3.2	欠損値への対応	40
3.2.1	欠損値が発生するメカニズム	40
3.2.2	欠損値への対応のフロー	45
3.2.3	欠損値の発生パターンの可視化	46
3.2.4	リストワイズ法	50
3.2.5	ペアワイズ法	51
3.2.6	平均値代入法	52
3.2.7	回帰代入法	55
3.2.8	確率的回帰代入法	56
3.2.9	完全情報最尤推定法	57
3.2.10	多重代入法	59
3.3	外れ値の検出	61
3.3.1	外れ値とは	61
3.3.2	外れ値検出のアプローチ	63
3.3.3	統計モデル	64
3.3.4	データの空間的な近さに基づくモデル	67
3.3.5	高次元の外れ値	71
3.4	連続データの離散化	76
3.4.1	等間隔区間による離散化	78
3.4.2	等頻度区間による離散化	78
3.4.3	カイマージ	79
3.4.4	情報エントロピーを用いた離散化	80
3.4.5	最小記述長原理を用いた離散化	81
3.5	属性選択	81
3.5.1	属性選択の手法の分類	82
3.5.2	フィルタ法のアルゴリズム	82
3.5.3	相関に基づく属性選択	83
3.5.4	情報量に基づく属性選択	83
3.5.5	データの近さに基づく属性選択	84
第4章	パターンの発見	86
4.1	予測モデルの構築	86
4.1.1	機械学習による予測モデル構築	87
4.1.2	予測モデル構築のプロセス	94
4.1.3	予測問題の設定	95
4.1.4	特徴量の構築	97
4.1.5	ハイパーパラメータの最適化	97
4.1.6	予測モデルの構築	105
4.1.7	予測モデル構築・評価における属性選択	110

4.1.8	不均衡データへの対応	111
4.1.9	並列計算による高速化	118
4.1.10	複数の予測モデルの結合	120
4.1.11	実データに対する分析：顧客の解約予測	124
4.1.12	実データに対する分析：顧客の購買予測	140
4.2	頻出パターンの抽出	145
4.2.1	頻出パターンマイニング	145
4.2.2	実データに対する分析：POSデータの頻出パターンマイニング	151
4.2.3	冗長性の低いパターンの抽出	156
4.2.4	系列パターンマイニング	161
4.2.5	実データに対する分析：POSデータの系列パターンマイニング	166
第5章	データ分析の例	169
5.1	StudentLife Studyの概要	169
5.2	データの理解	170
5.2.1	データの取得	170
5.2.2	データの概要	171
5.2.3	センサーデータ	171
5.2.4	購買履歴データ	172
5.3	分析計画の立案	172
5.3.1	予測問題の設定	172
5.3.2	飲食品の購買を予測するために使用する特徴量の検討	173
5.3.3	構築した予測モデルの活用方法の検討	174
5.4	データの加工	175
5.4.1	入退館時刻の推定	177
5.4.2	建物内での購買有無の判定	179
5.4.3	建物内の会話時間・回数の算出	181
5.4.4	建物内のアクティブ割合の算出	181
5.5	予測モデルの構築・評価	182
5.5.1	特徴量の作成	182
5.5.2	訓練期間とテスト期間の定義	185
5.5.3	予測モデルの構築	186
付録A	主な予測アルゴリズムの概要	188
A.1	決定木	188
A.1.1	アルゴリズムの概要	188
A.1.2	Rでの実行	188
A.2	ランダムフォレスト	191
A.2.1	不均衡データ	192

x 目次

A.2.2	Rでの実行	192
A.3	サポートベクタマシン	192
A.3.1	アルゴリズムの定式化	192
A.3.2	クラスウェイトの調整	194
A.3.3	クラス確率の推定	195
A.3.4	Rでの実行	196
付録B	caretパッケージで利用できるアルゴリズム	197
付録C	ELKIの使用方法	204
参考文献		210
索引		215