본 저자는 이 책을 쓸 때에 가능한 한 쉽게 설명하려고 노력하였습니다. 모델링 기법에 대한 이론 및 모델 구축 과정에 대한 설명을 간단한 예제로부터 시작해서 복잡한 예제로 끝을 맺고 있습니다. 데이터 애널리틱스를 전공하는 독자는 모든 예제를 실습해봄으로써 자신의 지식을 심화시킬 수 있고, 데이터 애널리틱스를 처음 접하는 독자는 간단한 예제를 실습해봄으로써 데이터 애널리틱스에 쉽게 입문할 수 있습니다.

이 책의 집필에서부터 출간에 이르기까지 오랜 시간이 걸렸습니다. 그 동안에 관련 소프트웨어의 새로운 버전들이 나와서 책의 내용을 수정하거나 추가해야 하는 어려움이 있었지만 마침내 책을 출간하게 되었습니다. 이 책의 출간에 도움을 주셨던 분들께 감사드립니다.

데이터를 분석하여 새로운 지식을 도출하는 과정은, 그 결과가 성공이었든 실패이었든 간에, 언제나 저에게 즐거움을 주었습니다. 이 책을 통해서 독자 여러분들도 그 즐거움을 느껴보시기 바랍니다.

2020년 7월 15일
이 재 식

01부 기본 개념

01장 / 서론

1. 인공지능, 머신러닝, 딥러닝 3
2. 데이터 사이언스와 데이터 애널리틱스 4
3. 데이터 마이닝과 데이터 애널리틱스 6
4. 이 책의 구성 7
5. 참고문헌 8

02장 / 데이터 마이닝

1. 데이터, 정보, 지식 10
2. 데이터의 속성 11
   - 2.1 범주형 속성 12
   - 2.2 수치형 속성 12
   - 2.3 기타 속성 13
3. 데이터 마이닝의 탄생 13
4. 데이터 마이닝의 정의 14
   - 4.1 비즈니스 프로세스 14
   - 4.2 대량의 데이터 14
   - 4.3 의미 있는 패턴과 규칙 15
5. 데이터 마이닝의 유형 15
   - 5.1 가설 검정 15
   - 5.2 방향성 데이터 마이닝 15
   - 5.3 무방향성 데이터 마이닝 15
6. 데이터 마이닝의 단계 16
   - 6.1 KDD2.0 16
   - 6.2 CRISP-DM 28
   - 6.3 두 방법론의 비교 34
7. 참고문헌 35

03장 머신러닝

1. 문제를 푸는 방법 36
2. 머신러닝의 정의 37
3. 머신러닝의 유형 38
- 3.1 지도 학습 38
- 3.2 비지도 학습 39
- 3.3 준지도 학습 39
- 3.4 강화 학습 40

4. 머신러닝의 기법들 40
- 4.1 지도 학습 기법 40
- 4.2 비지도 학습 기법 41
- 4.3 준지도 학습 기법 42
- 4.4 강화 학습 기법 43

5. 참고문헌 45

04장 모델 구축

1. 모델의 정의 46
2. 모델의 구축 과정 47
- 2.1 훈련 데이터 집합의 용도 50
- 2.2 과대적합의 발생 50
- 2.3 검증 데이터 집합의 용도 51
- 2.4 테스트 데이터 집합의 용도 51
- 2.5 스코어 데이터 집합의 용도 52

3. 편향되지 않은 모델의 구축 52
- 3.1 반복적 무작위 서브샘플링 검증 방법 53
- 3.2 K-폴드 교차 검증 방법 53
- 3.3 단일 관측값 제거 교차 검증 방법 54
- 3.4 부트스트랩 방법 55

4. 모델의 평가 56
- 4.1 회귀 평가 척도 57
- 4.2 분류 평가 척도 58

5. 편향과 편차 간의 상충 관계 65
- 5.1 훈련 데이터 집합의 평균으로 만든 모델 67
- 5.2 선형 회귀 모델 68
- 5.3 2차 다항 회귀 모델 69
- 5.4 6차 다항 회귀 모델 70
- 5.5 네 개 모델의 비교 71

6. 과대적합의 방지 또는 제거 72
- 6.1 속성 선정 73
- 6.2 균등화 76
- 6.3 조기 종료 76
- 6.4 드롭아웃과 배치 정규화 76
- 6.5 가지치기 77
- 6.6 앙상블 방법 77

7. 모델 데이터 집합 77
- 7.1 모델 데이터 집합의 크기와 밀도 77
- 7.2 오버샘플링 79
- 7.3 결측값 81

8. 모델의 비교 및 선정 83
- 8.1 ROC 곡선의 비교 83
- 8.2 통계적 검정에 의한 비교 84

9. 앙상블 방법에 의한 모델의 성능 향상 87
- 9.1 앙상블 방법 87
- 9.2 배깅 방법 89
- 9.3 부스팅 방법 90

10. 참고문헌 98

02부 _ 통계학 기반 기법

05장 / 회귀 분석

1. 회귀 분석 101
2. 단순 회귀 분석 103
   - 2.1 $\hat{\beta}$ 값 구하기: 최소자승법 104
   - 2.2 결정계수 R^2 105
   - 2.3 단순 회귀 분석의 예제 106
3. 다중 회귀 분석 110
   - 3.1 다중 회귀 분석의 예제 111
   - 3.2 단계별 회귀 분석 115
   - 3.3 조정된 결정계수 118
4. 다중 회귀 분석을 이용한 자동차 연비 추정 120
   - 4.1 A_Model: 모든 변수를 사용한 모델 123
   - 4.2 F_Model: 전방향 선택으로 선택한 변수를 사용한 모델 124
   - 4.3 B_Model: 역방향 제거로 선택한 변수를 사용한 모델 126
   - 4.4 S_Model: 양방향 선택과 제거로 선택한 변수를 사용한 모델 128
   - 4.5 P_Model: $\Pr(>|t|)$가 유의한 변수를 사용한 모델 130
   - 4.6 최종 모델의 선정 132
5. 참고문헌 133

06장 / 로지스틱 회귀 분석

1. 로지스틱 회귀 분석 134
2. 이진형 문제의 선형 회귀 모델 134
3. 이진형 문제의 로지스틱 회귀 모델 136
4. 로지스틱 회귀 분석을 이용한 잡지 구독 예측 141
   - 4.1 A_loModel: 모든 변수를 사용한 모델 143
   - 4.2 S_loModel: 양방향 선택과 제거로 선택한 변수를 사용한 모델 145

   - 4.3 P_loModel: $\Pr(>|z|)$가 유의한 변수를 사용한 모델 147
   - 4.4 최종 모델의 선정 148
5. 참고문헌 150

07장 / 선형 판별 분석

1. 선형 판별 분석 151
2. 선형 판별 분석 방법 151
   - 2.1 중심과의 거리를 이용하는 방법 152
   - 2.2 회귀를 이용하는 방법 154
   - 2.3 피셔의 선형 판별 방법 157
3. 선형 판별 분석을 이용한 대출 결정 162
4. 참고문헌 166

08장 / 베이즈 분류기

1. 베이즈 분류기 167
2. 베이즈 정리 167
3. 베이즈 분류기의 이해 169
4. 베이즈 분류기의 예제 171
   - 4.1 수치형 변수가 없는 경우: 예제 8.1 171
   - 4.2 구매 여부 개수가 0일 경우 175
   - 4.3 수치형 변수가 있는 경우: 예제 8.2 176
5. 베이즈 분류기를 이용한 스팸 메일 판정 180
6. 참고문헌 186

03부 _ 머신러닝 기반 기법

09장 / 군집 분석

1. 군집 분석 188
2. 군집의 의미 189
3. 근접성 191
   3.1 수치형 속성 191
   3.2 범주형 속성 191
   3.3 군집 간의 거리 측정 193
4. 클러스터링 결과의 평가 척도 194
5. 클러스터링을 위한 데이터 준비 195
   5.1 속성값 조정 195
   5.2 가중치 부여 196
6. 계층적 클러스터링: 예제 9.1 196
   6.1 병합적 클러스터링 196
   6.2 분할적 클러스터링 205
   6.3 클러스터링 결과의 평가 211
7. K-평균 클러스터링 214
   7.1 K-평균 클러스터링의 단계 214
   7.2 K-평균 클러스터링: 예제 9.2 221
   7.3 초기 무작위 중심의 선택 226
   7.4 K값의 설정 230
8. K-평균 클러스터링을 이용한 피교육자 군집 분석 231
9. 참고문헌 237

10장 / 연관 분석

1. 연관 분석 238
2. 연관 규칙 239
3. 연관 규칙의 도출 과정 241
   3.1 아이템의 상세화 수준 결정 241
   3.2 거래 데이터로부터 아이템집합 생성 242
   3.3 아이템집합이 판매된 거래 건수와 확률 산출 243
   3.4 아이템집합 가지치기 244
   3.5 연관 규칙 생성 247
   3.6 생성된 연관 규칙 평가 249
4. 연관 규칙 도출 연습 251
5. 연관 분석을 이용한 시장바구니 분석 253
6. 순차 패턴 분석 258
7. 유용한 순차 패턴의 발견 260
8. 순차 패턴 분석을 이용한 제품 구매 순서 분석 263
9. 참고문헌 266

11장 / 의사결정 트리

1. 의사결정 트리 267
2. 의사결정 트리의 용도 267
   2.1 분류 267
   2.2 점수 부여 270
   2.3 추산 271
3. 의사결정 트리의 형태 272
4. 의사결정 트리의 구축 273
   4.1 기본 과정 273
   4.2 분지 273
   4.3 의사결정 트리의 구축 단계 275
   4.4 의사결정 트리의 평가 276
   4.5 의사결정 트리에서 규칙의 추출 276

5. 최상 분지 속성의 선정 277
5.1 분지 속성 선정의 중요성 277
5.2 최상 분지 속성의 선정 기준 278
5.3 최상 분지 속성의 선정 과정 280
6. 의사결정 트리 구축 과정의 예제 282
6.1 엔트로피 분지 방법 282
6.2 지니 분지 방법 288
7. 의사결정 트리의 가지치기 292
7.1 가지치기의 필요성 292
7.2 오류 감소 가지치기 294
8. 의사결정 트리를 이용한 개인 신용 평가 299
9. 참고문헌 307

12장 랜덤 포레스트
1. 랜덤 트리 308
2. 랜덤 포레스트 309
3. 랜덤 포레스트를 이용한 고객 이탈 예측 315
4. 참고문헌 324

13장 새로운 회귀 분석 기법들
1. 균등화된 회귀 분석 325
1.1 균등화 325
1.2 균등화된 회귀 분석의 유형 327
1.3 엑셀을 사용한 균등화된 회귀 분석 329
2. 균등화된 회귀 분석을 이용한 자동차 연비 추정 337
3. $\hat{\beta}$ 값 구하기: 기울기 하강법 347

4. 기울기 하강법을 이용한 회귀 모델 구축 356
4.1 R을 사용한 기울기 하강법 358
4.2 텐서플로를 사용한 기울기 하강법 360
5. 참고문헌 364

14장 서포트 벡터 머신
1. 서포트 벡터 머신 367
2. 서포트 벡터 머신의 이해 367
3. 서포트 벡터 머신의 최적화 문제 수식화 370
4. 엑셀을 사용한 서포트 벡터 머신: 예제 14.1 374
5. 선형 분리 불가능 문제: 여유 변수의 도입 379
5.1 여유 변수를 도입한 SVM의 최적화 문제 수식 380
5.2 엑셀을 사용한 SVM: 예제 14.2 381
5.3 균등화 파라미터 386
6. 선형 분리 불가능 문제: 커널 트릭의 사용 388
6.1 엑셀을 사용한 SVM: 예제 14.3 390
6.2 커널 함수 392
6.3 엑셀을 사용한 SVM: 예제 14.4 395
7. 서포트 벡터 머신을 이용한 유방암 판정 397
8. 참고문헌 403

15장 인공 신경망
1. 인공 신경망 404
2. 인공 신경망의 구성 404
2.1 처리 요소 404
2.2 처리 요소의 입력과 출력 405
2.3 처리 요소의 결합과 계층의 결합 405

2.4 가중치와 활성 함수 406
2.5 학습 기능 407

3. 역전파 알고리즘 407
3.1 전방향 단계 407
3.2 역방향 단계 408
3.3 역전파 알고리즘의 과정: 예제 15.1 408
3.4 가중치 수정의 빈도 411

4. 활성 함수 412
5. 비선형 분류 415
6. XOR 문제를 푸는 인공 신경망: 예제 15.2 417
7. 범주형 속성의 인코딩 418
7.1 N개-중-1개 인코딩 419
7.2 N개-중-M개 인코딩 419
7.3 온도계 인코딩 421

8. 인공 신경망을 이용한 심장질환 판정 422
9. 참고문헌 432

16장 / 딥러닝

1. 딥러닝의 개요 433
1.1 기울기 소실 현상 433
1.2 기울기 소실 현상의 극복 437
1.3 과대적합의 방지 443

2. 심층 신경망 445
2.1 심층 신경망을 이용한 동물 유형 판정 445

3. 합성곱 신경망 454
3.1 합성곱 계층 454
3.2 풀링 계층 457

3.3 합성곱 신경망의 차원 계산 459
3.4 합성곱 신경망을 이용한 필기체 숫자 판독 462

4. 순환 신경망 467
4.1 순환 신경망 구조의 유형 470
4.2 초기 순환 신경망 모델의 단점 극복 471
4.3 순환 신경망을 이용한 문장 예측 478
4.4 순환 신경망을 이용한 감성 분석 482

5. 참고문헌 489

부록 A. R 설치하기 491
부록 B. R Studio 설치하기 494
부록 C. Anaconda, R, R Studio,Tensorflow 설치하기 497
부록 D. R 환경의 Tensorflow 2.0 버전 코드 506

찾아보기 513