1부 하둡을 활용한 데이터 과학의 개요 020

1장 데이터 과학 021

1.1 **데이터 과학이란 무엇인가? 023**
1.2 **데이터 과학의 예: 검색 광고 024**
1.3 **데이터 과학의 간략한 역사 025**
  1.3.1 통계학과 머신 러닝 026
  1.3.2 인터넷 거인들이 가져온 혁신 027
  1.3.3 현대 기업의 데이터 과학 028
1.4 **데이터 과학자가 되는 길 029**
  1.4.1 데이터 엔지니어 029
  1.4.2 응용과학자 030
  1.4.3 데이터 과학자로 전직하는 방법 031
  1.4.4 데이터 과학자가 갖춰야 할 소프트 스킬 033
1.5 **데이터 과학팀 구성하기 034**
1.6 **데이터 과학 프로젝트의 생명 주기 035**
  1.6.1 적절한 질문 036
  1.6.2 데이터 입수 037
  1.6.3 데이터 정제: 데이터 품질 관리 038
  1.6.4 데이터 탐색과 모델 특징 변수 설계 039
  1.6.5 모델 구축과 튜닝 040
  1.6.6 운영 시스템에 배포 041
1.7 **데이터 과학 프로젝트의 관리 041**
1.8 **요약 043**

2장 데이터 과학의 활용 사례 045

2.1 **빅데이터: 변화의 원동력 047**
  2.1.1 규모: 더 많은 데이터의 활용 047
  2.1.2 다양성: 더 많은 데이터 유형 048
  2.1.3 속도: 더 빠른 데이터 유입 049
2.2 **비즈니스 활용 사례 049**
  2.2.1 제품 추천 049
  2.2.2 고객 이탈 분석 051
  2.2.3 고객 세분화 051

2.2.4 영업 리드 우선순위 결정 052
2.2.5 감성 분석 054
2.2.6 이상 거래 탐지 055
2.2.7 유지 보수 예측 055
2.2.8 장바구니 분석 056
2.2.9 데이터 기반 의료 진단 057
2.2.10 환자의 재입원 예측 058
2.2.11 변칙 접근 탐지 059
2.2.12 보험 위험 분석 059
2.2.13 유정/가스정의 생산량 예측 060

2.3 요약 **060**

3장 하둡과 데이터 과학 061

3.1 하둡이란 무엇인가? **063**
3.1.1 분산 파일 시스템 063
3.1.2 리소스 관리자와 스케줄러 066
3.1.3 분산 데이터 처리 프레임워크 067

3.2 하둡의 진화 과정 **072**

3.3 데이터 과학용 하둡 도구 **074**
3.3.1 아파치 스쿱 074
3.3.2 아파치 플럼 075
3.3.3 아파치 하이브 075
3.3.4 아파치 피그 078
3.3.5 아파치 스파크 079
3.3.6 R 081
3.3.7 파이썬 083
3.3.8 자바 머신 러닝 패키지 084

3.4 하둡이 데이터 과학자에게 유용한 이유 **084**
3.4.1 저비용 스토리지 085
3.4.2 스키마 온 리드 086
3.4.3 비정형 데이터와 반정형 데이터 087
3.4.4 다양한 언어 지원 087
3.4.5 견고한 스케줄링과 리소스 관리 088
3.4.6 분산 시스템 추상화 레벨 089
3.4.7 대규모 데이터에 기반한 모델 구축 091
3.4.8 대규모 데이터에 모델을 적용 092

3.5 요약 **093**

2부 하둡을 활용한 데이터 준비와 시각화 094

4장 하둡을 활용한 데이터 입수 095

4.1 **하둡 데이터 레이크 097**
4.2 **HDFS 099**
4.3 **파일을 HDFS로 직접 전송하기 100**
4.4 **파일을 하이브 테이블로 가져오기 101**
  4.4.1 CSV 파일을 하이브 테이블로 가져오기 102
4.5 **스파크를 사용해 데이터를 하이브 테이블로 가져오기 106**
  4.5.1 스파크를 사용해 CSV 파일을 하이브로 가져오기 107
  4.5.2 스파크를 사용해 JSON 파일을 하이브로 가져오기 110
4.6 **아파치 스쿱을 활용한 관계형 데이터 입수 111**
  4.6.1 스쿱을 활용한 데이터 가져오기와 내보내기 111
  4.6.2 아파치 스쿱의 버전별 차이 113
  4.6.3 스쿱 버전 1을 사용한 기본 예제 114
4.7 **아파치 플럼을 활용한 데이터 스트림 입수 123**
  4.7.1 플럼을 활용한 웹 로그 수집 예제 126
4.8 **아파치 우지를 활용한 하둡 작업 및 데이터 흐름 관리 129**
4.9 **아파치 팔콘 132**
4.10 **새로운 데이터 입수 도구 134**
4.11 **요약 134**

5장 하둡을 활용한 데이터 개조 135

5.1 **하둡이 데이터 개조 작업에 필요한 이유 138**
5.2 **데이터 품질 138**
  5.2.1 데이터 품질이란 무엇인가? 138
  5.2.2 데이터 품질 이슈 다루기 140
  5.2.3 하둡을 사용한 데이터 품질 관리 145
5.3 **특징 행렬 147**
  5.3.1 적절한 특징 변수 선택하기 148
  5.3.2 샘플링: 인스턴스 선택 149
  5.3.3 특징 변수 생성 151
  5.3.4 텍스트 특징 변수 153
  5.3.5 시계열 특징 변수 159

  5.3.6 복잡한 데이터 유형에서 추출한 특징 변수 160
  5.3.7 특징 변수 조작 162
  5.3.8 차원 축소 163

5.4 **요약 167**

6장 데이터 탐색과 시각화 169

6.1 **왘 데이터를 시각화하는가? 171**
  6.1.1 동기 부여 예제: 네트워크 처리량 시각화하기 171
  6.1.2 애당초 없었던 혁신을 시각화하기 175

6.2 **데이터 차트 생성 176**
  6.2.1 비교 차트 178
  6.2.2 구성 차트 179
  6.2.3 분포 차트 182
  6.2.4 관계 차트 184

6.3 **데이터 과학에서 사용하는 시각화 차트 186**

6.4 **시각화 도구 187**
  6.4.1 R 187
  6.4.2 파이썬: matplotlib, seaborn 등 188
  6.4.3 SAS 188
  6.4.4 MATLAB 189
  6.4.5 Julia 189
  6.4.6 기타 시각화 도구 190

6.5 **하둡을 활용한 빅데이터 시각화 190**

6.6 **요약 191**

3부 하둡을 활용한 데이터 모델링 192

7장 하둡을 활용한 머신 러닝 193

7.1 **머신 러닝 개요 195**

7.2 **머신 러닝 용어 196**

7.3 **머신 러닝 작업 유형 197**

7.4 **빅데이터와 머신 러닝 198**

7.5 **머신 러닝 도구 199**

7.6 머신 러닝과 인공지능의 미래 201
7.7 요약 201

8장 예측 모델링 203

8.1 예측 모델링 개요 205
8.2 분류 vs 회귀 206
8.3 예측 모델 평가 208
    8.3.1 분류 모델 평가 209
    8.3.2 회귀 모델 평가 213
    8.3.3 교차 검증 213
8.4 지도 학습 알고리즘 214
8.5 빅데이터를 활용한 예측 모델 솔루션 구축 217
    8.5.1 모델 학습 217
    8.5.2 일괄 예측 219
    8.5.3 실시간 예측 220
8.6 예제: 감성 분석 221
    8.6.1 트윗 데이터셋 221
    8.6.2 데이터 준비하기 222
    8.6.3 특징 변수 생성 223
    8.6.4 분류 모델 구축 226
8.7 요약 228

9장 군집화 229

9.1 군집화 개요 231
9.2 군집화 활용 232
9.3 유사도 측정 방식 설계 233
    9.3.1 거리 함수 233
    9.3.2 유사도 함수 235
9.4 군집화 알고리즘 236
9.5 군집화 알고리즘의 예 237
    9.5.1 *k*-평균 군집화 237
    9.5.2 잠재 디리클레 할당 238
9.6 군집 평가와 군집 개수 선택 240

9.7 **빅데이터를 활용한 군집화 솔루션 구축 241**
9.8 **예제: LDA를 활용한 주제 모델링 244**
  9.8.1 데이터 입수 244
  9.8.2 특징 변수 생성 245
  9.8.3 LDA 실행 247
9.9 **요약 249**

10장 하둡을 활용한 이상 탐지 251

10.1 **이상 탐지 개요 253**
10.2 **이상 탐지 활용 253**
10.3 **데이터 내 이상 현상 유형 254**
10.4 **이상 탐지 기법 256**
  10.4.1 규칙 기반 기법 256
  10.4.2 지도 학습 기법 256
  10.4.3 비지도 학습 기법 258
  10.4.4 준지도 학습 기법 260
10.5 **이상 탐지 시스템 튜닝 260**
10.6 **하둡을 활용한 빅데이터 기반 이상 탐지 솔루션 구축 261**
10.7 **예제: 네트워크 침입 탐지 263**
  10.7.1 데이터 입수하기 266
  10.7.2 분류 모델 학습하기 268
  10.7.3 성능 평가하기 271
10.8 **요약 272**

11장 자연어 처리 273

11.1 **자연어 처리 275**
  11.1.1 자연어 처리의 역사 275
  11.1.2 자연어 처리의 활용 사례 276
  11.1.3 텍스트 분할 277
  11.1.4 품사 태깅 277
  11.1.5 개체명 인식 278
  11.1.6 감성 분석 278
  11.1.7 주제 모델링 279

11.2 **하둡의 자연어 처리 도구 279**
    11.2.1 스몰 모델 NLP 279
    11.2.2 빅 모델 NLP 281
11.3 **텍스트 표현 모델 282**
    11.3.1 Bag-of-Words 283
    11.3.2 Word2Vec 284
11.4 **감성 분석 예제 285**
    11.4.1 스탠포드 CoreNLP 285
    11.4.2 스파크를 활용한 감성 분석 285
11.5 **요약 291**

12장 하둡과 데이터 과학의 미래 293

12.1 **자동 데이터 탐색 295**
12.2 **딥러닝 297**
12.3 **요약 300**

부록 301

A.1 HDFS 퀵스타트 **302**
    A.1.1 퀵 명령 303
A.2 **참고 자료 309**
    A.2.1 하둡과 스파크에 관한 일반적인 정보 309
    A.2.2 하둡과 스파크 설치 레시피 310
    A.2.3 HDFS 311
    A.2.4 맵리듀스 311
    A.2.5 스파크 311
    A.2.6 필수 도구 312
    A.2.7 머신 러닝 312

찾아보기 314