데이터 과학자와 데이터 엔지니어를 위한 인터뷰 문답집

데이터 과학자와 데이터 엔지니어를 위한 인터뷰 문답집

1쇄 발행 2020년 6월 30일

지은이 Hulu 데이터 과학팀
옮긴이 김태헌
펴낸이 장성두
펴낸곳 제이펍

출판신고 2009년 11월 10일 제406-2009-000087호
주소 경기도 파주시 회동길 159 3층 3-B호
전화 070-8201-9010 / **팩스** 02-6280-0405
홈페이지 www.jpub.kr / **원고투고** jeipub@gmail.com
독자문의 readers.jpub@gmail.com / **교재문의** jeipubmarketer@gmail.com

편집팀 이종무, 이민숙, 최병찬, 이주원 / **소통·기획팀** 민지환, 송찬수, 강민철 / **회계팀** 김유미
진행 및 교정·교열 장성두 / **내지디자인** 이민숙 / **표지디자인** 미디어픽스
용지 에스에이치페이퍼 / **인쇄** 한승인쇄 / **제본** 광우제책사

ISBN 979-11-90665-23-0 (93000)
값 34,000원

데이터 과학자와 데이터 엔지니어를 위한 인터뷰 문답집

The Quest for Machine Learning

주거웨 책임편집 / **Hulu 데이터 과학팀** 지음 / **김태헌** 옮김

Jpub 제이펍

드리는 말씀

- 이 책에 기재된 내용을 기반으로 한 운용 결과에 대해 저자, 역자, 소프트웨어 개발자 및 제공자, 제이펍 출판사는 일체의 책임을 지지 않음으로 양해 바랍니다.
- 이 책에 등장하는 각 회사명, 제품명은 일반적으로 각 회사의 등록 상표 또는 상표입니다.
  본문 중에는 ™, ©, ® 마크 등이 표시되어 있지 않습니다.
- 이 책에서 소개한 URL 등은 시간이 지나면 변경될 수 있습니다.
- 용어는 바이두백과사전(百度百科)의 영문 표기, 《기계학습》(오일석 저), 대학수학회와 한국통계학회의 용어집을 참고하였습니다.
- '데이터 과학자/데이터 엔지니어를 위한 스킬 트리(기술 로드맵)' PDF 파일은 www.jpub.kr의 이 책 소개 페이지에서 다운로드하실 수 있습니다.
- 책의 내용과 관련된 문의 사항은 지은이나 출판사로 연락해 주시기 바랍니다.
  - 옮긴이: https://data-manyo.tistory.com
  - 출판사: readers.jpub@gmail.com

차 례

추천사 xvii
머리말 xx
옮긴이 머리말 xxiii
프롤로그 xxv
베타리더 후기 xxxvii

CHAPTER 1 피처 엔지니어링 1

1 피처 정규화 3

질문 수치형 데이터에 대한 피처 정규화가 중요한 이유는 무엇인가? 난이도 ★☆☆☆☆ 3

2 범주형 피처 6

질문 데이터 정제 작업을 진행할 때 범주형 피처는 어떻게 처리해야 할까요? 난이도 ★★☆☆☆ 6

3 고차원 결합 피처의 처리 방법 9

질문 결합 피처란 무엇일까요? 고차원 결합 피처는 어떤 방식으로 피처 엔지니어링 해야 할까요? 난이도 ★★☆☆☆ 9

4 결합 피처 12

질문 효율적인 결합 피처는 어떻게 찾을 수 있을까요? 난이도 ★★☆☆☆ 12

5 텍스트 표현 모델 14

질문 텍스트 표현 모델에는 어떤 것들이 있나요? 각 모델의 장단점은 무엇인가요? 난이도 ★★☆☆☆ 14

6 Word2Vec 17

질문 Word2Vec은 어떤 알고리즘이고, LDA와는 어떤 차이점과 관련이 있을까요? 난이도 ★★★☆☆ 17

7 이미지 데이터가 부족할 때는 어떻게 처리해야 할까요? 20

질문 이미지 분류 문제에서 훈련 데이터가 부족하다면 어떤 문제를 일으킬까요? 어떻게 데이터 부족이 야기하는 문제들을 완화할 수 있을까요? 난이도 ★★☆☆☆ 20

CHAPTER 2 모델 평가 23

1 평가 지표의 한계 25

질문 1 정확도의 한계성 난이도 ★☆☆☆☆ 25

질문 2 정밀도와 재현율의 균형 난이도 ★☆☆☆☆ 27

질문 3 평균제곱근오차의 예외 난이도 ★☆☆☆☆ 29

2 ROC 곡선 31

질문 1 ROC 곡선이란 무엇일까요? 난이도 ★☆☆☆☆ 31

질문 2 ROC 곡선은 어떻게 그릴까요? 난이도 ★★☆☆☆ 32

질문 3 AUC는 어떻게 계산할까요? 난이도 ★★☆☆☆ 35

질문 4 ROC 곡선과 P-R 곡선을 비교해 보세요. 난이도 ★★★☆☆ 35

잠시 쉬어가기... | ROC 곡선의 유래 37

3 코사인 거리의 응용 38

질문 1 어떤 상황에서 유클리드 거리 대신 코사인 유사도를 사용하는지를 학습과 연구 경험을 토대로 설명해 보세요. 난이도 ★★☆☆☆ 38

질문 2 코사인 거리는 엄격한 의미에서의 거리가 맞습니까? 난이도 ★★★☆☆ 40

4 A/B 테스트의 함정 43

질문 1 모델에 대해 충분한 오프라인 평가를 진행하더라도 왜 온라인에서 다시 한번 A/B 테스트를 진행해야 하는 것일까요? 난이도 ★☆☆☆☆ 43

질문 2 A/B 테스트는 어떻게 진행해야 하나요? 난이도 ★☆☆☆☆ 44

질문 3 실험군과 대조군은 어떻게 분류할까요? 난이도 ★★☆☆☆ 44

5 모델 평가 방법 46

질문 1 모델 평가 과정에서 사용할 수 있는 검증 방법에는 어떤 것들이 있고, 각 방법의 장단점에 관해 설명해 주세요. 난이도 ★★☆☆☆ 46

질문 2 부트스트래핑 과정 중에서 n개의 샘플에 대해 n번의 샘플링을 하는데, n이 무한대로 커진다면 한 번도 추출되지 않는 데이터의 수는 얼마나 될까요? 난이도 ★★★☆☆ 48

6 하이퍼파라미터 튜닝 49

질문 하이퍼파라미터 최적화 방법에 대해 설명해 주세요. 난이도 ★★★☆☆ 49

잠시 쉬어가기... | 구글은 하이퍼파라미터 최적화 알고리즘으로 더 맛있는 쿠키를 굽는다 51

7 과적합과 과소적합 52

질문 1 모델 평가 과정에서 과적합과 과소적합이란 어떤 현상을 뜻하는 것일까요? 난이도 ★☆☆☆☆ 52

질문 2 과적합과 과소적합의 위험을 낮출 수 있는 몇 가지 방법에 대해 설명해 주세요. 난이도 ★★☆☆☆ 53

CHAPTER 3 클래식 알고리즘 55

1 서포트 벡터 머신 57

질문 1 공간상에서 선형분리 가능한 두 점이 각각 SVM이 분리한 초평면상으로 투영된다면, 이 점들의 초평면상의 투영도 선형분리가 가능한가요? 난이도 ★★★☆☆ 59

질문 2 SVM 훈련오차를 0이 되도록 만드는 파라미터 세트가 존재할까요? 난이도 ★★★☆☆ 62

질문 3 훈련오차가 0인 SVM 분류기는 반드시 존재할까요? 난이도 ★★★★☆ 64

질문 4 여유 변수를 추가한 SVM의 훈련오차는 0이 될 수 있나요? 난이도 ★★★☆☆ 65

잠시 쉬어가기... | SVM의 창시자 블라디미르 배프닉과 그의 유명 동료들 65

2 로지스틱 회귀 67

질문 1 선형회귀와 비교했을 때 로지스틱 회귀의 다른 점은 무엇이 있을까요? 난이도 ★★☆☆☆ 67

질문 2 로지스틱 회귀를 사용하여 다중분류 문제를 해결할 때 자주 사용되는 방법은 어떤 것이 있을까요? 각각은 어떤 상황에서 쓰이고, 어떤 관계를 맺고 있을까요? 난이도 ★★★☆☆ 69

3 의사결정 트리 71

질문 1 의사결정 트리에서 사용하는 휴리스틱 함수에는 어떤 것들이 있나요? 난이도 ★★☆☆☆ 72

질문 2 의사결정 트리에 대한 가지치기는 어떻게 진행할까요? 난이도 ★★★☆☆ 79

잠시 쉬어가기... | 오컴의 면도날(Occam's Razor 혹은 Ockham's Razor) 84

CHAPTER 4 차원축소 85

1 PCA 최대분산 이론 87

질문 주성분은 어떻게 정의할까요? 이 정의에서 출발하여 주성분을 추출하는 목적을 달성하기 위해 어떻게 목적함수를 설계해야 할까요? 해당 목적함수로 어떻게 PCA 문제의 해를 구할 수 있을까요? 난이도 ★★☆☆☆ 87

2 PCA 최소제곱오차 이론 92

질문 PCA의 해는 사실 최적의 투영 방향, 즉 하나의 직선을 찾는 것인데, 이는 선형회귀 문제의 목적과 일치합니다. 그렇다면 회귀의 시각에서 PCA의 목적을 정의하고 문제의 해를 구하는 방법이 있을까요? 난이도 ★★☆☆☆ 92

3 선형판별분석 96

질문 레이블이 있는 데이터의 차원축소 과정에서 레이블 정보를 잃지 않게 하려면 어떤 목적함수를 설정해야 할까요? 그리고 해당 목표를 달성하기 위해 어떻게 해를 구해야 할까요? 난이도 ★★☆☆☆ 97

4 선형판별분석과 주성분분석 101

질문 LDA와 PCA는 전통적인 차원축소 알고리즘입니다. 응용적 관점에서 이들 원리의 같고 다름을 분석해 보세요. 수학적 관점과 목적함수에 대해 두 차원축소 알고리즘은 어떤 차이점과 연관성이 있는지 설명해 보세요. 난이도 ★★☆☆☆ 101

CHAPTER 5 비지도학습 107

1 k평균 클러스터링 109

질문 1 K평균 알고리즘이 작동하는 방법에 대해 구체적으로 설명해 주세요. 난이도 ★★☆☆☆ 110

질문 2 K평균 알고리즘의 장단점은 무엇인가요? 알고리즘을 튜닝하는 방법에는 어떤 것들이 있나요? 난이도 ★★★☆☆ 112

질문 3 K평균 알고리즘의 단점을 개선한 모델은 어떤 것들이 있을까요? 난이도 ★★★☆☆ 115

질문 4 K평균 알고리즘의 수렴성에 대해 증명하세요. 난이도 ★★★★☆ 117

2 가우스 혼합 모델 121

질문 가우스 혼합 모델의 핵심 아이디어는 무엇인가요? 알고리즘은 어떻게 반복적으로 계산되나요? 난이도 ★★☆☆☆ 122

3 자기 조직화 지도 125

질문 1 SOM 알고리즘에 대해 설명해 보세요.
SOM과 K평균 알고리즘은 어떤 차이점이 있나요? 난이도 ★★★☆☆ 125

질문 2 SOM은 어떻게 설계해야 할까요? 그리고 네트워크
훈련 파라미터는 어떻게 설정해야 할까요? 난이도 ★★★☆☆ 129

4 클러스터링 알고리즘 평가 131

질문 외부 라벨(정답) 데이터가 없다고 가정한다면 어떻게
두 클러스터링 알고리즘을 비교할 수 있을까요? 난이도 ★★★☆☆ 131

CHAPTER 6 확률 그래프 모델 137

1 확률 그래프 모델의 결합확률분포 139

질문 1 그림 6.1(a)에서의 베이지안 네트워크의
결합확률분포 식을 작성해 주세요. 난이도 ★☆☆☆☆ 139

질문 2 그림 6.1(b)에서의 마르코프 네트워크의
결합확률분포 식을 작성해 주세요. 난이도 ★☆☆☆☆ 140

2 확률 그래프 표현 142

질문 1 나이브 베이즈 모델의 원리에 관해 설명하고
확률 그래프 모델로 표현해 보세요. 난이도 ★★☆☆☆ 142

질문 2 최대 엔트로피 모델의 원리에 관해 설명하고
확률 그래프 모델로 표현해 보세요. 난이도 ★★☆☆☆ 143

3 생성모델과 판별모델 146

질문 자주 보이는 확률 그래프 모델 중에는
어떤 생성모델과 판별모델이 있나요? 난이도 ★★★☆☆ 146

4 마르코프 모델 148

질문 1 단어 분할 문제에서 은닉 마르코프 모델을 사용하여
모델링하고 훈련하는 방법에 대해 설명해 주세요. 난이도 ★★★☆☆ 149

질문 2 최대 엔트로피 마르코프 모델에서 레이블
편향 문제가 생기는 이유는 무엇일까요?
이 문제에 대한 해결 방안은 무엇인가요? 난이도 ★★★★☆ 151

잠시 쉬어가기... | 베이지안 이론과 '하나님의 존재' 154

5 토픽 모델 156

질문 1 자주 사용하는 토픽 모델에는 어떤 것들이 있나요?
토픽 모델과 그 원리에 대해 설명해 주세요. 난이도 ★★☆☆☆ 156

질문 2 LDA 모델의 토픽 개수는 어떻게 정해야 할까요? 난이도 ★★☆☆☆ 159

질문 3 어떻게 토픽 모델을 사용하여 추천 시스템의
콜드 스타트 문제를 풀 수 있나요? 난이도 ★★★☆☆ 161

CHAPTER 7 최적화 알고리즘 163

1 지도학습에서의 손실함수 165

질문 지도학습법에서 사용하는 손실함수는 어떤 것들이 있나요? 예제와 함께 설명하고 각 손실함수의 특징도 함께 말해 주세요. 난이도 ★☆☆☆☆ 165

2 머신러닝에서의 최적화 문제 169

질문 머신러닝에서의 최적화 문제에서 어떤 컨벡스 최적화 문제와 넌컨벡스 최적화 문제가 존재할까요? 예를 들어 설명해 보세요. 난이도 ★★☆☆☆ 169

3 전통적인 최적화 알고리즘 172

질문 제약 조건이 없는 최적화 문제에서의 최적화 방법에는 어떤 것들이 있을까요? 난이도 ★★☆☆☆ 172

잠시 쉬어가기... | 빠른 역제곱근(Fast Inverse Square Root) 175

4 경사하강법 검증 방법 177

질문 목적함수 기울기의 계산에 대한 검증은 어떻게 진행할까요? 난이도 ★★★☆☆ 177

5 확률적 경사하강법 180

질문 훈련 데이터 크기가 매우 큰 상황에서 전통적인 경사하강법을 사용한다면 어떤 문제가 생길까요? 이에 대한 개선 방안은 무엇인가요? 난이도 ★☆☆☆☆ 180

잠시 쉬어가기... | 기울기 연산자 ∇의 발음 183

6 확률적 경사하강법의 가속 184

질문 1 확률적 경사하강법이 효과를 상실하게 되는 원인 — 돌을 더듬어 가며 산을 내려오다. 난이도 ★★☆☆☆ 184

질문 2 해결 방법 — 관성 보존과 환경 감지 난이도 ★★★☆☆ 187

7 L1 정규화와 희소성 192

질문 L1 정규화를 사용해 모델 파라미터에 희소성을 갖게 할 수 있는 원리는 무엇인가요? 난이도 ★★★☆☆ 193

CHAPTER 8 샘플링 199

1 샘플링의 역할 201

질문 머신러닝에서 샘플링이 어떻게 활용되는지 설명해 보세요. 난이도 ★★☆☆☆ 201

2 균등분포의 난수 204

질문 어떻게 프로그래밍을 통해 균등분포 난수 생성기를 만들 수 있을까요? 난이도 ★☆☆☆☆ 204

3 자주 사용하는 샘플링 방법 207

질문 특정한 분포에 대해 설계된 샘플링 방법 외에 알고 있는 샘플링 방법이나 샘플링 전략에는 어떤 것들이 있는지, 그리고 그들의 주요 아이디어와 구체적인 진행 과정을 설명해 주세요. 난이도 ★★★☆☆ 207

4 가우스 분포 샘플링 212

질문 가우스 분포에서의 샘플링에 대해 설명해 주세요. 난이도 ★★★☆☆ 212

잠시 쉬어가기... | 정규분포를 왜 가우스 분포라고도 할까? 217

5 마르코프 체인 몬테카를로 219

질문1 MCMC 샘플링 방법의 주요 아이디어에 대해 설명해 보세요. 난이도 ★☆☆☆☆ 219

질문2 자주 사용하는 MCMC 샘플링 방법 몇 가지를 소개해 주세요. 난이도 ★★☆☆☆ 220

질문3 MCMC 샘플링은 어떻게 상호 독립적인 샘플을 얻을 수 있을까요? 난이도 ★★☆☆☆ 222

잠시 쉬어가기... | MCMC 샘플링을 활용한 암호 해석 223

6 베이지안 네트워크 샘플링 225

질문 베이지안 네트워크 샘플링 과정을 설명해 주세요. 만약 일부 변수의 주변분포를 고려해야 한다면 어떻게 샘플링해야 할까요? 만약 네트워크에 관측변수가 포함되어 있다면 또 어떻게 샘플링해야 할까요? 난이도 ★★★☆☆ 226

7 불균형 샘플 집합에서의 리샘플링 230

질문 이진분류 문제에 대해 훈련 세트 중에 양성-음성 샘플 비율이 불균형할 때 어떻게 데이터를 처리해야 더 좋은 분류 모델을 훈련할 수 있을까요? 난이도 ★★★☆☆ 230

CHAPTER 9 피드 포워드 신경망 235

1 다층 퍼셉트론과 부울 함수 237

질문 1 다층 퍼셉트론으로 XOR 논리를 표현하려면 최소 몇 개의 은닉층이 필요할까요? 난이도 ★★☆☆☆ 237

질문 2 하나의 은닉층만 사용하여 n차원 입력을 가진 임의의 부울함수를 구현하기 위해서는 몇 개의 은닉 노드가 필요할까요? 난이도 ★★★☆☆ 240

질문 3 다수의 은닉층이 있는 상황을 고려했을 때, n차원 입력을 포함하는 임의의 부울함수는 최소 몇 개의 네트워크 노드와 네트워크 층을 필요로 할까요? 난이도 ★★★☆☆ 243

2 딥러닝의 활성화 함수 245

질문 1 자주 사용하는 활성화 함수와 해당 활성화 함수의 도함수를 작성해 주세요. 난이도 ★☆☆☆☆ 245

질문 2 왜 시그모이드와 Tanh 활성화 함수는 그래디언트 소실 현상을 일으킬까요? 난이도 ★★☆☆☆ 246

질문 3 ReLU 계열의 활성화 함수는 시그모이드, Tanh 활성화 함수와 비교했을 때 어떤 장점이 있나요? 이들의 한계는 무엇이며, 어떤 개선 방안들이 있나요? 난이도 ★★★☆☆ 247

3 다층 퍼셉트론의 오차역전파 알고리즘 249

질문 1 다층 퍼셉트론의 제곱오차와 크로스 엔트로피 손실함수에 대해 작성해 보세요. 난이도 ★★☆☆☆ 250

질문 2 문제 1에서 정의한 손실함수에 기반하여 각 층 파라미터가 업데이트하는 그래디언트 계산 공식을 유도하세요. 난이도 ★★★★☆ 251

질문 3 제곱오차 손실함수와 크로스 엔트로피 손실함수는 각각 어떤 상황에서 사용해야 할까요? 난이도 ★★★☆☆ 253

잠시 쉬어가기... | 신경망의 흥망성쇠 254

4 딥러닝 훈련 테크닉 257

질문 1 신경망을 훈련할 때 모든 파라미터를 0으로 초기화해도 될까요? 난이도 ★☆☆☆☆ 257

질문 2 과적합을 방지할 수 있는 드롭아웃의 원리와 구현 방법을 말해 주세요. 난이도 ★★★☆☆ 258

질문 3 배치 정규화의 주된 동기와 원리는 무엇인가요? 합성곱 신경망에서는 어떻게 사용되나요? 난이도 ★★★☆☆ 260

5 합성곱 신경망 263

질문 1 컨볼루션 작업의 본질적인 특성에는 희소 상호작용과 파라미터 공유가 있습니다. 이 두 특성에 대해 설명해 보세요. 난이도 ★★☆☆☆ 264

질문 2 자주 사용하는 풀링 방법에는 어떤 것들이 있나요? 풀링은 어떤 작용을 하나요? 난이도 ★★★☆☆ 266
질문 3 합성곱 신경망은 텍스트 분류 문제에서 어떻게 사용되고 있나요? 난이도 ★★★☆☆ 268
6 ResNet 271
질문 ResNet이 나오게 된 배경과 핵심 이론은 무엇인가요? 난이도 ★★★☆☆ 272
잠시 쉬어가기... | 제프리 힌튼의 전설적인 삶 275

CHAPTER 10 순환신경망 277

1 순환신경망과 합성곱 신경망 279
질문 텍스트 데이터를 다룰 때 순환신경망과 피드 포워드 신경망은 각각 어떤 특징이 있나요? 난이도 ★☆☆☆☆ 279
2 순환신경망의 그래디언트 소실 문제 281
질문 순환신경망에서 그래디언트 소실이나 그래디언트 폭발이 일어나는 원인은 무엇일까요? 어떤 개선 방안들이 있나요? 난이도 ★★☆☆☆ 281
3 순환신경망의 활성화 함수 284
질문 순환신경망에서 ReLU를 활성화 함수로 사용해도 될까요? 난이도 ★★★☆☆ 284
4 LSTM 네트워크 286
질문 1 LSTM은 어떻게 장단기 기억 기능을 구현할 수 있나요? 난이도 ★★☆☆☆ 286
질문 2 LSTM의 각 모듈은 어떤 활성화 함수를 사용하고 있나요? 다른 활성화 함수를 사용해도 될까요? 난이도 ★★★☆☆ 288
5 Seq2Seq 모델 290
질문 1 Seq2Seq 모델은 무엇인가요? Seq2Seq 모델은 어떤 장점이 있나요? 난이도 ★★☆☆☆ 290
질문 2 Seq2Seq 모델에서 디코딩할 때 자주 사용하는 방법들은 어떤 것들이 있나요? 난이도 ★★★☆☆ 292
6 어텐션 메커니즘 294
질문 Seq2Seq 모델의 어떤 문제를 해결하기 위해 어텐션 메커니즘을 도입했나요? 왜 양방향 순환신경망 모델을 사용할까요? 난이도 ★★★★☆ 294
잠시 쉬어가기... | 벤지오 형제 298

CHAPTER 11 강화학습 299

1 강화학습 기초 301

질문 1 강화학습에는 어떤 기본 개념들이 있나요? 마리오 보물찾기 문제에서 이러한 개념을 어떻게 정의할 수 있을까요? 난이도 ★☆☆☆☆ 301

질문 2 그림 11.1에서 주어진 마리오와 보물의 위치에서 가치 반복을 고려해 어떻게 하면 최적의 길을 찾을 수 있을지 설명해 주세요. 난이도 ★★☆☆☆ 303

질문 3 그림 11.1에서 주어진 마리오와 보물의 위치에서 정책 반복을 고려해 어떻게 하면 최적의 길을 찾을 수 있을지 설명해 주세요. 난이도 ★★☆☆☆ 305

2 비디오 게임에서의 강화학습 308

질문 심층강화학습이란 무엇일까요? 전통적인 강화학습과 어떤 점이 다른가요? 난이도 ★★★☆☆ 308

잠시 쉬어가기... | 도파민에서 강화학습까지 311

3 폴리시 그래디언트 313

질문 폴리시 그래디언트는 무엇인가요? 전통적인 Q-러닝과 어떤 차이점이 있으며, Q-러닝 대비 어떤 장점이 있나요? 난이도 ★★★★☆ 313

4 탐색과 이용 317

질문 에인전트가 환경과 상호작용을 하는 과정에서 탐색과 이용이란 무엇을 말하는 것일까요? 어떻게 탐색과 이용 사이에 균형을 맞출 수 있을까요? 난이도 ★★★☆☆ 317

CHAPTER 12 앙상블 학습 323

1 앙상블 학습의 종류 325

질문 앙상블 학습에는 어떤 종류가 있나요? 이들 사이에는 어떤 공통점 혹은 차이점이 있나요? 난이도 ★☆☆☆☆ 325

잠시 쉬어가기... | 레오 브레이만 328

2 앙상블 학습 단계와 예제 329

질문 앙상블 학습에는 어떤 기본적인 단계가 있나요? 앙상블 학습 예제를 통해 설명해 주세요. 난이도 ★★☆☆☆ 329

3 기초 분류기 332

질문 1 자주 사용하는 기초 분류기는 무엇인가요? 난이도 ★☆☆☆☆ 332

질문 2 랜덤 포레스트의 기초 분류기를 의사결정 트리에서 선형분류기 혹은 K-최근접 이웃 알고리즘으로 대체할 수 있을까요? 난이도 ★★☆☆☆ 333

4 편향과 분산 334

질문 1 편향과 분산이란 무엇일까요? 난이도 ★★☆☆☆ 334

질문 2 '편향과 분산 줄이기' 관점에서 부스팅과 배깅의 원리를 설명해 주세요. 난이도 ★★★☆☆ 336

5 GBDT 알고리즘의 기본 원리 338

질문 1 GBDT의 기본 원리는 무엇일까요? 난이도 ★★☆☆☆ 338

질문 2 그래디언트 부스팅과 경사하강법 사이에는 어떤 연관성과 차이점이 존재할까요? 난이도 ★★☆☆☆ 340

질문 3 GBDT의 장점과 한계에는 어떤 것들이 있을까요? 난이도 ★★☆☆☆ 341

6 XGBoost와 GBDT의 차이점, 그리고 연관성 342

질문 XGBoost와 GBDT의 차이점, 그리고 연관성에는 어떤 것들이 있나요? 난이도 ★★★☆☆ 342

잠시 쉬어가기... | 머신러닝 경진대회 캐글 345

CHAPTER 13 생성적 적대 신경망 347

1 처음 만나는 GANs의 비밀 349

질문 1 GANs의 기본 아이디어와 훈련 프로세스에 대해 설명해 주세요. 난이도 ★☆☆☆☆ 349

질문 2 GANs의 가치함수 난이도 ★★★☆☆ 352

질문 3 GANs에서는 어떻게 대량 확률추론 계산을 피할 수 있을까요? 난이도 ★★☆☆☆ 354

질문 4 GANs 실제 훈련 중에 만날 수 있는 문제들은 어떤 것들이 있을까요? 난이도 ★★★★☆ 355

2 WGAN: 저차원의 유령을 잡아라 357

질문 1 GANs의 함정: GANs에 존재하는 어떤 문제들이 모델 훈련 효과를 저하시켰을까요? 난이도 ★★★☆☆ 358

질문 2 WGAN은 위 문제를 어떤 방법으로 개선했나요? 와서스타인 거리란 무엇일까요? 난이도 ★★★★☆ 360

질문 3 어떻게 구체적으로 와서스타인 거리를 사용해 WGAN 알고리즘을 구현할 수 있을까요? 난이도 ★★★★★ 363

3 DCGAN: GANs이 합성곱을 만났을 때 365
질문 생성기와 판별기에서 어떻게 합성곱 구조를 설계할 수 있을까요? 난이도 ★★★☆☆ 367
4 ALI 372
질문 생성 네트워크와 추론 네트워크의 융합 난이도 ★★★☆☆ 374
5 IRGAN: 이산 샘플의 생성 377
질문 GANs을 사용해 음성 샘플을 생성하세요. 난이도 ★★★★★ 378
6 SeqGAN: 텍스트 시퀀스 생성 382
질문 1 텍스트로 구성된 시퀀스를 생성해 문장을 표현하려면 생성기를 어떻게 만들어야 할까요? 난이도 ★★☆☆☆ 383
질문 2 시퀀스 생성기를 훈련할 때 사용하는 최적화 목표는 일반적으로 무엇인가요? GANs 프레임과 어떤 차이점이 있나요? 난이도 ★★★★★ 384
질문 3 생성기의 최적화 목표가 있다면 어떻게 생성기 파라미터에 대한 그래디언트를 구할 수 있을까요? 난이도 ★★★★★ 388

CHAPTER 14 인공지능의 응용 현황 391

1 알고리즘 마케팅 393
2 게임에서의 인공지능 409
3 자율 주행에서의 AI 428
4 기계 번역 439
5 인간과 컴퓨터 상호작용 443

에필로그 및 저자 소개 449
참고문헌 465
찾아보기 470

칭화대학교 컴퓨터 공학과 동문인 주거웨와 그녀의 동료들이 출판한 《데이터 과학자와 데이터 엔지니어를 위한 인터뷰 문답집》의 추천사를 쓸 수 있어서 영광입니다.

두말할 필요 없이 인공지능은 만물이 소생하는 봄과 같은 부흥기에 접어들었는데, 그 열기는 마치 여름을 방불케 하기도 합니다. 하지만 저는 오히려 수확의 계절인 가을에 비유하고 싶습니다. 지금 세상을 휩쓸고 있는 인공지능의 물결은 사실 이삼십여 년에 걸친 이론과 알고리즘 연구들이 조금씩 누적된 결과물입니다. (당연히 빅데이터와 강해진 연산 능력을 갖춘 컴퓨터 덕도 있겠죠.) 하지만 지금은 본질적으로 대부분 '약한 인공지능' 범위에 들어갑니다. 이번 물결이 끝나면 다음 물결의 시발점은 '강한 인공지능'이 되길 바라지만, 이론적으로 매우 높은 난이도를 요구하고 있기 때문에 언제 '강한 인공지능'이 실현될지에 대해서는 아무도 알지 못합니다. 따라서 우리들은 이 기회를 잘 잡아 '약한 인공지능'이 후하게 주고 있는 풍성한 '열매'를 잘 거둬야 할 것입니다. 가지각색의 인공지능 응용 상품은 우리의 생활 속으로 들어와, 예전에 인터넷이 그랬던 것처럼 사회와 경제 전체에 영향을 주고 있습니다.

그러나 이러한 '열매'는 아무나 얻을 수 없습니다. '열매'를 거둘 자격이 되는 사람들만 풍작의 기쁨을 누리겠지요. 이런 사람들이 바로 일선에서 인공지능과 머신러닝을 연구하거나 일하며 새로운 알고리즘과 방법을 끊임없이 시도하고 있는 사람들(일반적으로 데이터 과학자나 데이터 엔지니어 등)일 것입니다. 이들은 요구를 이해하고, 데이터를 수집하며, 알고리즘을 설계하고, 반복해서 실험하고, 또 최적화가 완료될 때까지 끊임없이 노력합니다. 이들이 바로 새로운 인공지능 기술의 선구자이자 원동력입니다.

여러분은 이들 중 한 사람이 되고 싶은가요?
그렇다면 어떻게 해야 이른 시일 내에 이들 중 한 명이 될 수 있을까요?

아마도 이 책이 여러분을 한 발짝 나아가도록 도와줄 것입니다. 많은 우수한 연구원과 프로그래머가 인공지능과 머신러닝의 실전 응용문제들을 해결하기 위해 부지런히 일하고 있지만, 실제 문제들을 해결하기 위한 기술 서적은 부족한 실정입니다. 이 책이 이런 부족한 부분을 채워 줄 수 있을 것이라 기대합니다. 이 책은 간단한 내용부터 복잡한 내용까지 차례대로 전개되며, 머신러닝 각각의 영역을 포괄하는 간결한 문답 형식으로 되어 있습니다. 따라서 독자가 인공지능 분야에 종사하기 위해 알아야 할 기술을 잘 설명하고 있는 동시에 독자들이 필수 기술을 익힐 수 있도록 도와줍니다.

저는 주거웨를 꽤 오래전부터 알고 지냈습니다. 그녀는 저희 학과에서 유명한 '공부의 신'이었습니다. ACM SIGMOD* Test of Time Award를 받기도 했었죠. 귀국 후 그녀는 학과에서 열리는 활동에 종종 참가하곤 했습니다. 그녀와 함께 일하는 동료들도 모두 훌륭한 배경을 가진 분들이라고 알고 있습니다. 게다가, 이 책은 산업계에서 매일같이 머신러닝을 적용하고 있는 데이터 과학자들이 함께 만든 책이기 때문에 결코 여러분을 실망시키지 않을 것입니다.

많은 분이 이 책을 통해 더 멋진 데이터 과학자, 알고리즘 엔지니어, 인공지능 실무자로 거듭날 수 있기를 바랍니다. 제가 이끄는 연구진은 얼마 전 '구가九歌**'를 자동으로 만들어 내는 시스템을 개발했습니다. 2017년에는 CCTV의 〈사람을 이기는 기계 지능〉이란 프로그램에도 소개되었는데, 많은 사람이 실제 사람이 쓴 시와 기계가 만든 시를 구별하지 못하게 하는 정도까지 성공했습니다. 이 기술 역시 이 책에 나오는 LSTM과 Seq2Seq 모델을 적용한 것입니다.

순마오숭(孙茂松) / 칭화대학교 컴퓨터 공학과 교수

---

* 옮긴이 ACM SIGMOD는 데이터베이스 최고 권위의 학회 중 하나입니다.

** 옮긴이 '구가(九歌)'는 초나라 민간의 제사 노래이며, 음악, 가사, 무도가 혼합되어 이루어진 것입니다. 총 11편으로 구성되어 있는데, 여기서는 단순히 구가라는 시와 비슷한 구조를 가진 시를 만들어 내는 알고리즘이라고 생각하면 되겠습니다.

이 책은 주거웨 박사가 편집하고 15명의 Hulu 소속의 데이터 과학자가 함께 쓴 창의적이고 실용적인 면이 돋보이는 책입니다. 인공지능과 머신러닝에 대한 이해를 높여 소프트웨어 엔지니어와 데이터 과학자 모두를 AI 전문가로 거듭날 수 있도록 도와줄 것입니다. 동시에 데이터 과학자들을 훌륭한 AI 연구자들로 만들어 줄 것입니다.

해리 셤(Harry Shum) / 마이크로소프트 글로벌 수석부사장, IEEE 펠로우, ACM 펠로우

컴퓨터 이론과 알고리즘은 사람들에게 자주 냉대를 받습니다. 왜냐하면 그들과 실제 응용 사이를 이어주는 다리가 없기 때문입니다. 주거웨 박사와 그녀의 동료들이 쓴 이 책은 어떻게 그들을 잇는 다리를 만들어 줄 수 있는지에 대해 가르쳐 주고 있습니다. 이 책을 통해 컴퓨터 관련 종사자들은 이론적인 부분에서 크게 도약할 것이며, 비전공자 출신들도 컴퓨터 과학이란 위대한 도구를 더 잘 이해할 수 있을 것입니다.

우쥔(Wu Jun) / 《수학의 아름다움(數學之美)》, 《물결의 정점에서(浪潮之巔)》 저자

시장에 쏟아져 나오고 있는 머신러닝 관련 서적 중에서 Hulu 데이터 과학자들이 출판한 이 책은 매우 특별합니다. 이 책은 단순히 다른 사람들의 말을 옮기거나 학술적인 시각에서 머신러닝의 이론과 알고리즘 체계를 정리한 것이 아닙니다. 일선에서 일하고 있는 데이터 과학자들의 시각으로 인터뷰, 실전 모델링, 그리고 응용 사례들을 중점으로 머신러닝을 설명하고 있습니다. 그래서 데이터 과학자를 꿈꾸는 독자들에게는 더 빠르게 꿈을 이룰 수 있도록 도와줄 것입니다. 특히, 여러 명의 실전 전문가가 힘을 합쳐 만든 것임에도 내용이 상당히 체계적이라 더욱 독보적입니다.

리우펑(Liu Peng) / 《알고리즘 마케팅(计算广告)》 저자, iFLYTEK 부사장

머신러닝 데이터 과학자로서의 자기 수양

데이터 과학자로 향하는 고급 과정은 순탄치 않을 것입니다. 《선형대수》, 《통계학습 방법》, 《단단한 머신러닝》, 《패턴인식과 머신러닝》, 《심층 학습》, 《목 디스크 회복 가이드》와 같은 책들이 여러분의 회사 생활 내내 함께할 것입니다.

개인적으로 훌륭한 데이터 과학자가 되기 위해서는 관련 지식에 대한 체계적이고 완벽한 준비뿐만 아니라, 알고리즘 모델에 대하여 마음 깊숙한 곳에서부터 나오는 열정과 연구 작업에 대한 장인 정신이 있어야 한다고 생각합니다. 여기서 말하는 장인 정신이란, 문제를 발견하는 눈빛, 문제를 해결하는 탐구 정신, 그리고 집착에 가까운 끈기를 뜻합니다.

이런 장인 정신은 저희 팀원들의 생활 곳곳에서 엿볼 수 있습니다. 건물 아래로 내려가 점심을 먹기 위해 엘리베이터를 기다릴 때마다, 그리고 화장실을 가는 사이에 엘리베이터를 놓칠까 여러 모델을 세워 놓고 서로 다른 시간대에서 평균적으로 엘리베이터를 기다리는 시간을 계산해 최적의 타이밍을 계산합니다. 그리고 회사 앞 호수에 노을의 물결이 햇빛에 반짝이는 것을 보고 이러한 빛의 상태가 이미지 인식을 어렵게 만드는 이유에 대해서 고민합니다. 쇼핑 앱을 열어 셀 수 없이 많은 상품을 보며 어떤 추천 시스템을 만들어야 사용자들이 좋아할 만한 상품만을 추천할 수 있을까를 고민합니다. 연구에 대한 열정만 있다면 이런 사소한 일들도 머신러닝의 문을 여는 열쇠가 될 수 있습니다.

많은 데이터 과학자들이 일상생활 중 찰나의 순간에 영감을 얻어 상품화까지 진행하고는 합니다. 같은 팀에서 일하는 한 동료는 어떤 국내 애플리케이션으로 드라마를 보던 중 아주 간편하게 오픈 크레딧과 엔딩 크레딧을 건너뛸 수 있는 기능이 있는 것을 발견했습니다. 소비자 입장에서 이런 기능은 큰 편의를 가져다줄 수 있다고 생각해서 우리 플랫폼의 원천 데이터를 살펴보았으나 일부 영상만 오픈 크레딧과 엔딩 크레딧 정보가 있음을 알았고, 이러한 정보 모두를 사람이 수기로 태깅tagging했다는 것을 알아냈습니다. 백만 이상의 콘텐츠를 보유한 영상 회사에서 모든 콘텐츠에 태깅한다는 것은 터무니없는 생각일 것입니다. 결국, 그는 광범위한 연구와 계속되는 시도 끝에 딥러닝에 기반을 둔 엔딩 크레딧 자동 검출 모델을 개발해 냈습니다. 반복적이고 충분한 실험 끝에 모두가 만족할 만한 결과를 얻을 수 있었고, 미국에 발명 특허까지 신청하며 상품화의 길로 가고 있습니다.

알고리즘 연구를 비즈니스에 적용하는 것과 순수한 학술 연구 사이에는 큰 차이점이 존재하는데, 바로 사용자(유저)의 시각에서 문제를 생각하는 것입니다. 많은 경우에 데이터 과학자들이 만든 상품이 데이터 지표를 향상시키더라도 이것이 진정으로 사용자가 원하는 것인지는 다시 생각해 봐야 합니다. 이런 기준으로 많은 모델 중에 가장 적합한 모델을 선택하고, 빠른 도입과 반복적인 테스트를 통해 상품화라는 결과를 얻어야 합니다. 이러한 창조 정신과 도전 정신을 통해 '장인 정신'이 표현되는 것이라 생각합니다.

장인 정신도 물론 중요하지만, 관련 지식도 데이터 과학자로 성공하기 위한 필수 불가결한 기초입니다. 이것이 바로 저희가 이 책을 쓰게 된 이유이기도 합니다. 튼튼한 수학 기초, 알고리즘 시스템에 대한 완전한 이해, 모델에 대한 깊은 이해는 우리가 독자들에게 전달하고 싶은 정수입니다. 이 책 앞부분에 나오는 피처 엔지니어링(특성 공학), 모델 평가, 전통적인 모델 등은 머신러닝 영역의 기초이므로 반드시 자기 자신의 것으로 만들어야 합니다. 그리고 연구원 혹은 응용 영역의 전문가가 되고 싶다면 머신러닝 스킬 트리skill tree의 각 가지 중 특정 부분에 대한 깊은 지식을 키워 나가야 합니다.

많은 사람이 맥주와 기저귀에 관한 이야기를 알고 있을 것입니다. 하지만 비교적 완전하고 안정적인 추천 시스템을 구축하기 위해서는 차원축소(제4장)뿐만 아니라 최적화 알고리즘(제7장), 딥러닝(제9장, 제10장), 강화학습(제11장) 등에 대해서도 깊게 이해해야 합니다. 그리고 계속된 연구를 통해서만이 최신 학술 기술을 상품에 녹일 수 있을 것입니다. 예를 들어, 만약 스킬 트리에서 마르코프 모델과 토픽 모델(제6장)을 깊게 공부해 완전한 확률 그래프 모델 지식 네트워크를 만들고, 순환신경망(제10장)의 이론 체계를 접목하고 자신만의 이해와 생각을 더한다면 기계 번역, 음성 비서 등 자연어 처리 응용 영역에서 활약할 수 있을 것입니다.

데이터 과학자로 향하는 길은 순탄치 않을 것입니다. 하지만 그 길에는 아름다움과 광활함이 함께할 것입니다. 여러분이 해야 할 일은 자신이 어떤 일을 하고 싶은지를 명확히 하고, 묵묵히 이 책의 내용을 최대한 습득한 후, 조용히 이 책을 덮고서 생활 속 사소한 곳에서 머신러닝의 매력을 느껴보는 것입니다.

Hulu 데이터 과학팀 드림

데이터 과학은 (통계학, 컴퓨터 공학, 의료, 경제 등) 서로 다른 영역에서 온 사람들이 계속해서 데이터 과학이라는 '정의'를 확립해 가는, 동적dynamic인 융합 학문이라고 생각합니다. 따라서 기본적인 지식의 틀이 어느 정도 정해져 있는 개발자 인터뷰(면접)와는 달리 제대로 정리된 인터뷰 관련 서적을 찾기 어렵습니다. 데이터 과학은 인터뷰를 위해 (고급 통계 지식, 머신러닝/딥러닝 알고리즘, 컴퓨터 공학 지식, 도메인 지식 등) 기본적으로 준비해야 할 범위가 매우 넓을 수밖에 없는데, 지원자뿐만 아니라 현업에 종사하는 분들도 1시간 남짓의 짧은 시간 내에 후보자의 역량을 파악할 수 있는 '좋은 인터뷰 질문'을 던지기 위해 고심에 고심을 거듭하고 있습니다.

이 책을 접하자마자 데이터 과학자 인터뷰를 앞둔 지원자, 그리고 좋은 데이터 과학자를 채용해야 하는 현업 종사자들께 좋은 가이드라인이 될 수 있을 것 같아 한 치의 망설임도 없이 번역 의뢰를 했었습니다. 이 책은 인터뷰 준비뿐만 아니라 데이터 과학자가 갖춰야 할 소양(기본 지식)을 측정하는 도구로도 사용할 수 있을 것입니다. 데이터 과학 내에도 비전vision, 자연어 처리NLP, 마케팅, 비즈니스 데이터 분석 등 다양한 분야가 있어서 모든 분야의 지식을 섭렵하기란 매우 어려운 일인데, 만약 다른 분야에 대한 폭넓은 지식과 이해력을 키우고 싶은 독자라면 이 책의 일독을 권합니다. 특히, AI 관련 조직의 관리자급 자리에서 일하는 분들께 큰 도움이 될 것입니다.

이 책의 저자는 스탠퍼드대학교 출신의 주거웨 박사가 이끄는, 미국의 HULU라는 회사의 데이터 과학팀입니다. HULU는 넷플릭스Netflix의 급성장으로 위기를 느낀 월트 디즈니 컴퍼니, 21세기 폭스, 타임 워너 등 전통 미디어 대기업이 공동투자로 만

든 OTT 서비스입니다. 아직 한국에는 서비스되고 있지 않아 국내 독자들에게는 다소 생소한 이름일 것입니다. 하지만 약 3천만 유료 구독자를 가진, 이제는 추천 시스템의 대명사가 되어버린 넷플릭스처럼 추천 시스템, 알고리즘 마케팅, 영상 및 텍스트 분석에 특화된 AI 기술을 보유한 기업입니다. HULU팀이 현업에서 다루는 머신러닝 기술의 범위가 넓기 때문에 이 책에서 전통 머신러닝, 이미지 분석, 텍스트 분석 등 다양한 분야의 지식이 총 망라될 수 있었던 것 같다는 생각이 듭니다.

이번 책은 15명의 데이터 과학자가 공동으로 집필했기 때문에 각 장마다, 혹은 각 절마다 글의 스타일이 조금씩 달라 번역 과정에서 최대한 원문의 뜻을 유지하며 문체를 통일시키는 데 주안점을 두었습니다. 번역에 있어서 아직 부족한 부분이 많음에도 저의 첫 번역서인《단단한 머신러닝》에 대해 힘이 되는 피드백을 많이 주셨기 때문에, 이번 책에서 조금 더 자신감을 가지고 잘 마무리할 수 있었던 것 같습니다.

두 번째 번역서를 잘 마무리할 수 있게 체력을 허락해 주신 하나님께 감사드리며, 좋은 책이 나올 수 있도록 교정 과정에 직접 참여해 주시고 고생해 주신 장성두 대표님께 감사의 말씀을 전합니다. 그리고 사랑하는 아내 유리나와 아들 라온이, 딸 라엘이에게 항상 사랑한다는 말을 전하고 싶습니다.

마지막으로, 이 책이 인터뷰를 앞둔 학생, 직장인 분들의 귀한 시간을 아껴줄 수 있는 소중한 책이 되길 기원합니다.

김태헌 드림

인공지능, 그 세 번의 물결

2018년 초, 취업 시즌이 다가왔을 때 '데이터 과학자'와 '알고리즘 엔지니어'는 최고의 인기 직업이었습니다.

'인공지능', '머신러닝', '딥러닝', '모델링', 'CNN' 등과 같은 단어가 일반인들의 대화 중에도 많이 언급되기 시작했고, '데이터베이스 구조', '연결 리스트', '배열' 등이 소프트웨어 엔지니어들의 필수 스킬로 자리 잡았습니다.

인공지능 기술은 사회구조, 직장, 교육 등의 영역에 혁명적인 변화를 몰고 오고 있습니다. 향후 몇 년간은 인공지능 기술이 전면적으로 보편화되는 시기인 동시에 해당 기술을 가진 인재들이 가장 부족할 시기이기도 합니다. 따라서 우리는 이 책을 통해 인공지능과 머신러닝에 관심 있는 독자들에게 이 분야의 기본 기능을 더 깊게 이해시키고 싶고, 이미 어느 정도 기본기가 있는 독자들에게는 인공지능과 머신러닝의 고수가 될 수 있도록 돕고 싶습니다.

책을 시작하기에 앞서서 제가 이해하고 있는 머신러닝의 배경과 역사를 소개하고, 왜 지금이 머신러닝을 배우기에 좋은 시기인지에 관해 설명하고자 합니다.

나, 그리고 인공지능

저의 학부 전공은 인공지능입니다. 제가 대학교에 다닐 때 칭화대학교 컴퓨터 공학과에는 6개의 반이 있었는데, 반마다 각자의 전공이 정해져 있었습니다. 제가 속한

3반의 전공은 인공지능이었습니다. 덕분에 저는 학부 시절부터 인공지능 영역의 최신 기술을 접할 수 있었습니다. 제가 수강했던 인공지능 입문 수업을 담당하시던 분은 존경하는 린야오루이林尧瑞 교수님이었는데, 《人工智能导论(인공지능 개론)》의 저자이시기도 합니다. 저희는 이 수업을 '원숭이 바나나 따 먹기'라고 불렀는데, 수업 가장 첫 질문이 어떻게 지능을 가진 원숭이가 스스로 블록을 조립해 천장에 달린 바나나를 따 먹을 수 있는지에 대한 것이기 때문입니다.

당시 칭화대학교 학부생 과정은 5년이었는데, 소수의 학생은 4학년 때 대학원생 프로젝트에 참여할 수 있었고, 6년째에 석사학위를 취득할 수 있었습니다. 저는 그 소수의 행운아 중 한 명이었습니다. 덕분에 4학년 때 칭화대학교 인공지능 연구실에 들어가 장보张钹 교수님을 따라 간단한 연구를 진행할 수 있었습니다. 장 교수님과 대학원생들 사이에서 저는 인공지능에 관련된 첨단 지식을 배울 수 있었습니다.

스탠퍼드대학교에 갓 입학했을 무렵에는 20여 명 정도가 모이는 소규모 점심 강의 brown bag에 참여한 적이 있었습니다. 시작되고 절반 정도 지났을 무렵에 교실 문이 열리더니 존 매카시John McCarthy 교수님이 "이곳에 공짜 점심이 있다는 소리를 들었네."라고 크게 말하며 들어왔습니다. 그리고는 교실 앞으로 가서 샌드위치 몇 개를 집어 다시 왔던 길을 통해 성큼성큼 돌아갔습니다. 강의를 기획했던 교수님은 조금 당황하더니 우리를 향해 다시 말했습니다. "세계에서 가장 유명한 과학자가 교실로 들어와 여러분의 음식을 가져가는 곳, 스탠퍼드에 온 것을 환영한다!"

혹시 모를 수도 있는 독자들을 위해 언급하자면, '인공지능Artificial Intelligence'이라는 단어가 바로 존 매카시 교수님에게서 나왔습니다.

저는 학부 전공이 인공지능이었고, 줄곧 인공지능에 흥미가 있었기 때문에 스탠퍼드에 와서도 인공지능 과목인 CS140을 들었습니다. 당시 이 과목을 가르치던 분은 닐스 닐슨Nils Nilsson 교수님이었습니다. 그는 또 다른 인공지능의 창시자이자 세계적인 전문가입니다. 그의 명작 《The Quest for Artificial Intelligence》는 많은 연구자에게 인용되고 있습니다. 닐슨 교수의 수업은 정말 흥미로웠는데, 저는 오늘날까지 그때의 필기 노트를 보관하고 있을 정도입니다.

사실대로 말하면, 제가 젊었을 때는 이런 최고의 과학자들과 같은 교실에 있다는 것이 얼마나 행운인지 깨닫지 못했고, 인공지능이 이렇게까지 주목받는 기술이 될지 몰랐습니다. 이런 최고의 기술은, 처음에는 소수의 사람만 이해하고 그 가치를 알아보는 것 같습니다.

그러나 제 박사 논문은 인공지능에 대한 것이 아니라 데이터베이스와 데이터 마이닝에 관련된 것이었습니다. 지금 돌이켜 보면, 저와 인공지능, 그리고 인공지능 대가들과의 만남은 인공지능의 세 번의 물결과 관련이 있습니다. 처음 인공지능 물결을 일으킨 장본인이 바로 존 매카시 시대의 연구자들인데, 1950년대부터 컴퓨터 과학과 인공지능 이론의 기초를 마련했습니다. 두 번째 물결은 칭화대학교 시기에 시작되었는데, 연구자들은 특정 영역에서 인공지능의 가능성을 확인할 수 있었습니다. 예를 들면, 조립 기계나 로봇, 전문가 시스템 등이 있습니다. 그리고 최근에 빅데이터와 머신러닝에 기반한 인공지능이 다시 부흥하고 있는데, 많은 사람이 이를 인공지능의 세 번째 물결이라고 부릅니다.

● 인공지능의 3차 물결

먼저, 간단하게 이 책에서 나오는 개념을 정리해 보겠습니다.

인공지능Artificial Intelligence이란, 기계가 인간의 지능을 가질 수 있게 하는 기술을 말합니다. 이 기술의 목적은 기계가 사람과 같은 인지, 사고, 행동, 문제 해결을 할 수 있도록 만드는 데 있습니다. 인공지능은 매우 폭넓은 기술인데, 자연어 처리, 컴퓨터 비전, 로보틱스, 논리 규칙 등을 포함하고 있으며, 컴퓨터 과학의 하위 분야로 보는 사람도 있습니다. 컴퓨터 과학 외에도 심리학, 인지과학, 사회학 등 다양한 학문이 융합된 학문입니다.

머신러닝은 컴퓨터가 주변 환경을 관찰하고 교류하여 정보를 얻어 학습하며, 계속해서 자기를 업데이트하고 개선해 나가는 것을 뜻합니다. 여러분은 컴퓨터 프로그램이 어떻게 작동하는지 이해하고 있을 것입니다. 프로그램은, 예를 들면 '지도의 출력'과 같이 컴퓨터가 집행할 수 있는 특정한 지령입니다. 그렇다면 머신러닝과 우리에게 익숙한 프로그램과의 차이점은 무엇일까요? 바로, 머신러닝은 프로그래머가

작성한 것이 아니라는 점입니다. 머신러닝은 기계가 대량의 데이터로부터 학습하는 것을 뜻합니다.

간단하게 말해, 대부분의 머신러닝 알고리즘은 훈련training과 테스트test, 두 단계로 나눌 수 있습니다. 이 두 단계는 겹쳐서 진행되기도 합니다. 훈련에는 일반적으로 훈련 데이터가 필요하고, 기계에 이전 사람들의 경험을 알려 줍니다. 예를 들면, 어떤 것이 고양이이고 어떤 것이 개인지, 그리고 어떤 물체를 봤을 때 정지해야 하는지 등을 알려줍니다. 훈련 학습의 결과는 기계가 작성한 프로그램 혹은 저장된 데이터인 모델model입니다. 전체적으로 봤을 때 훈련은 지도학습supervised learning과 비지도학습unsupervised learning 두 분류로 나뉩니다. 지도학습은 정답을 알려주는 선생님이 있는 경우에 비유할 수 있고, 비지도학습은 독학에 비유할 수 있는데, 기계 스스로 데이터에서 패턴과 특징을 찾는 것을 뜻합니다. 딥러닝deep learning은 머신러닝의 한 종류입니다. 주로 신경망을 토대로 만들어지며, 음성, 영상, 언어 이해 등 각 방면에서 활용됩니다.

먼저, 우리는 인공지능의 세 차례 물결에 대해 간단히 살펴보겠습니다. 각 물결(혹은 도약기)에는 어떤 특징이 있고, 이들은 어떻게 서로 다를까요? 그리고 서로 어떤 연관이 있을까요?

첫 번째 인공지능 물결은 대략 1950년대에 일어났습니다. 1956년, 다트머스 인공지능 연구 토론회에서 존 매카시에 의해 '인공지능'이란 개념이 정식으로 생겨났고, 현대 인공지능 학문의 기원으로 공식적인 인정을 받고 있습니다. 그리고 매카시와 MIT의 마빈 민스크Marvin Minsky는 '인공지능의 아버지'라는 칭호를 얻었습니다.

컴퓨터 발명 초기에 많은 컴퓨터 과학자가 인간이 발명해 낸 기계가 인류와 어떤 근본적인 차이가 있을까에 대해 진지하게 사고하고 토론했습니다. 튜링 기계Turing Machine와 튜링 테스트Turing Test는 바로 이러한 사고의 전형적인 결과물입니다. 최초의 인공지능 전문가들은 사상과 이론 측면에서 이미 많이 앞서 나갔는데, 초기부터 컴퓨터의 잠재력을 알아본 셈입니다. 우리가 지금 묻는 이런 문제는 대부분 그들이 이미 오래전에 사고하고 토론했던 질문들입니다. 예를 들어, '추론reasoning'은 무엇이고, 기계는 어떻게 추론하는가? '이해understanding'는 무엇이고, 기계는 어떻게 이해

하는가? 지식knowledge이란 무엇이고, 기계는 어떻게 지식을 얻고 표현할까? 언제쯤 기계는 사람과 구별하기 힘들 정도의 지능을 가질 수 있을까? 등의 질문이 있습니다. 이 시기에 많은 기초 이론이 생겨났으며, 인공지능의 기초 이론 이외에도 컴퓨터 공학, 컴퓨터 과학의 기초를 다지는 시기이기도 했습니다.

기술적인 측면에서 이야기하자면, 인공지능이 처음으로 크게 발전했던 시기였으며, 주로 논리에 기반을 둔 발전이었습니다. 1958년 매카시는 논리 언어인 LISP를 고안했습니다. 1950년대부터 1980년대까지 연구자들은 컴퓨터로 게임을 할 수 있고, 어느 정도는 자연어 이해까지 가능하다는 사실을 증명했습니다. 실험실에서는 로봇이 논리적인 판단을 하고, 나무 탑을 쌓았으며, 로봇 쥐가 미로를 탐색하며 스스로 장애물을 판단하기도 했습니다. 그리고 작은 자동차는 제약적인 환경에서 기초적인 자율 주행이 가능했습니다. 연구자들은 간단한 언어 이해와 물체 식별을 할 수 있는 신경망을 발명하기도 했습니다.

그러나 인공지능 발전 초기 이삼십 년 사이에 풍성한 연구 결과가 쏟아져 나왔음에도 실제 생활에 적용되지는 못했습니다. 1980년대 초, 인공지능은 응용 분야가 부족하다는 이유로 '빙하기'에 접어들게 됩니다. 제가 갓 대학교에 입학했을 무렵인 1980년대 말에서 1990년대 초까지 인공지능 과학자들은 새로운 길을 개척하기 시작했는데, '보편적인 지능 문제의 해결'에서 '특정 영역의 단일 문제 해결'로 연구의 목적을 바꿨습니다. '전문가 시스템'이란 개념이 탄생했고, 이는 처음으로 인공지능 연구 결과의 상업화를 위한 가능성을 제시했습니다.

컴퓨터 기술은 30년 정도의 발전을 겪으며 데이터베이스와 응용 방면에서 기초가 다져졌습니다. 연구자들은 인공지능이 데이터와 결합할 수 있는 가능성을 확인했는데, 가장 잘 결합할 수 있는 응용 형태가 바로 '전문가 시스템'이었습니다. 만약 심장병 관련 데이터와 같은 특정 영역의 데이터를 기계에 주입하고 일정한 논리를 가르쳐 준다면, 해당 기계는 '심장병 전문가'가 될 수 있을 것입니다.

병을 진단하거나 날씨를 예측하는 등 각 분야의 전문가 시스템은 실현 가능해 보이는 동시에 수요도 존재했기 때문에 당시 학술계에는 다시 한번 인공지능 열풍이 불기 시작했습니다. 하지만 이러한 전문가 시스템을 활용해 병 진단을 하려 할 때 문제

가 되는 것은 '어떻게 진단할지'에 대한 부분보다는 당시 대부분의 데이터가 디지털화되어 있지 않았다는 사실이었습니다. 당시 대부분의 의료 데이터는 수기로 작성되어 서류 형태로 보관되어 있었습니다. 설령, 어떤 분야의 데이터가 디지털화되어 있다 하더라도 서로 연결되어 있지 않은 로컬 컴퓨터에 저장되어 있어 사용하기 힘들었습니다.

따라서 전문가 시스템을 구축하려던 사람들은 되려 더 기초적인 작업에 몰두할 수밖에 없었습니다. 여기서 기초적인 작업이란, 간단히 말해 세상의 모든 정보를 디지털화하고 정량화하는 것입니다.

연구자들이 세상에 존재하는 책, 지도, 처방전 등을 디지털화하던 무렵에 인터넷이 등장해 광범위하게 사용되기 시작했고, 인터넷은 이러한 대규모 정보를 서로 이어주기 시작해 '빅데이터'라는 개념이 생겨났습니다. 동시에, 무어의 법칙Moore's law(마이크로칩의 성능이 2년마다 두 배로 증가한다는 경험적 예측)이 실제로 작용하며 컴퓨터 성능도 빠르게 발전했습니다. 컴퓨터 성능 향상에 따라 실험실이나 제한된 환경에서만 적용되던 인공지능 응용과 실제 생활 사이의 거리가 점점 좁혀졌습니다. 1997년에 딥블루DeepBlue는 세계 체스 챔피언 가리 카스파로프Garry Kasparov에게 승리를 거두었는데, 2017년 알파고AlphaGo가 바둑에서 이세돌 9단에게 이긴 것처럼 인공지능 역사에서 하나의 마일스톤으로 자리 잡았습니다. 컴퓨팅 파워가 향상되면서 목적이 특정한 단일 분야에서 기계가 사람을 이기는 일은 단지 시간 문제로 여겨졌습니다.

세 번째 인공지능 물결은 다른 두 기술 영역의 발전을 바탕으로 한 것인데, 하나는 컴퓨터 연산 능력이고, 다른 하나는 대규모 데이터입니다. 컴퓨팅 능력은 하드웨어, 분산 시스템, 클라우드 컴퓨팅 기술 등의 발전으로부터 온 것입니다. 최근에는 신경망을 위해 제작된 하드웨어 시스템neural-network-based computing이 인공지능 소프트웨어와 하드웨어의 결합이라는 대약진을 이뤄냈습니다. 대규모의 데이터는 몇십 년간 진행된 데이터 디지털화에 대한 노력과 인터넷의 발전 덕분에 진일보할 수 있었습니다. 예를 들어, 2001년에 런칭한 GPS 시스템은 사상 초유의 위치 데이터를 만들어 냈고, 스마트폰은 유례없는 양의 생활 데이터를 만들어 냈습니다. 뛰어난 연산력을 가진 컴퓨터의 탄생과 빅데이터의 결합은 머신러닝 알고리즘의 비약적인 발전을

촉진했습니다.

이번 인공지능 물결이 시작된 지 10여 년이 흘렀는데, 기술의 비약적인 발전은 유례없는 응용 분야의 발전까지 가져왔습니다. 최근의 인공지능 물결과 이전 두 차례 물결의 기본적인 차이점은 보편적인 응용 범위가 넓어지고 일반인의 생활 속까지 큰 영향을 준다는 점입니다. 달리 말하자면, 인공지능은 실험실을 떠나 우리들의 생활 속으로 들어오게 된 것입니다.

인공지능은 인간의 지능에 가까워질 수 있을까?

왜 이번 인공지능의 물결이 거세게 느껴질까요? 인공지능이 정말로 인간의 능력을 뛰어넘는 것일까요? 현재 인공지능 기술의 발전은 어떤 단계까지 왔을까요? 우리는 먼저 세 가지 팩트에 대해 살펴보겠습니다.

먼저, 인공지능은 역사상 처음으로 많은 복잡한 문제의 해결 방면에서 인간을 뛰어넘거나 곧 인간을 뛰어넘을 만한 능력을 갖췄습니다. 예를 들면, 이미지 인식, 영상 콘텐츠 해석, 기계 번역, 자율 주행, 바둑 등이 있습니다. 이러한 문제들은 인간에게도 어렵지 않은 문제이자 사람에 의해 해결되던 문제였습니다. 따라서 인공지능이 인류를 대체한다는 제목의 헤드라인이 신문을 장식하기 시작했습니다.

사실 단일 기술 방면에서, 특히 계산 관련 기술에서 이미 오래전부터 기계가 사람의 능력을 뛰어넘었고, 이런 기술이 광범위하게 응용되고 있습니다. 예를 들면, 내비게이션, 검색, 이미지 검색, 주가 거래 등이 있습니다. 이미 많은 사람이 음성을 통해 간단한 지령을 내리는 일에 익숙합니다. 그러나 상대적으로 이런 단순한 기술은 주로 '하나의 임무'를 완성하는 것에 국한되어 있고, 컴퓨터는 인간의 감지, 사고, 복잡한 판단, 감정 등에 대해서는 크게 관여하고 있지 못합니다.

그러나 최근 몇 년간 기계가 보여준 문제 해결 능력은 그 복잡성과 형식을 고려했을 때 이미 사람에 많이 가까워졌습니다. 예를 들면, 머신러닝에 기반을 둔 자율 주행 능력은 이미 성숙 단계에 접어들었는데, 이 기술은 인간의 생활 방식에 혁명적인 영향을 미칠 뿐만 아니라 도시 건설, 개인 소비 등에 광범위한 영향을 미칠 것입니다.

아마도 사람들은 더 이상 차를 소유하지 않아도 되거나, 차를 운전하는 방법조차 잊을지도 모릅니다. 사람들은 이 기술이 빠르게 현실화되어 가는 것을 바라보며 흥분하는 동시에 두려움을 가지고 있습니다. 기술이 가져올 편리함은 기대되지만, 너무 빠른 변화에 속수무책일까 걱정하는 것입니다.

이 외에도 컴퓨터의 자기 학습 능력은 계속해서 강해지고 있습니다. 현대 머신러닝 알고리즘, 특히 딥러닝 부류의 머신러닝 알고리즘의 발전은 기계의 행위가 더 이상 예측 가능한 '정도'나 '논리'를 뛰어넘어 사람이 이해하기 힘든 '블랙박스' 방식의 사고 능력을 갖추게 만듭니다.

그러나 자세히 살펴보면, 인공지능은 다양한 특수 영역에서 비약적인 발전을 보여주고 있지만, 아직 첫 번째 물결이 일어난 시기에 인공지능 선구자들이 이야기했던 범용적인 인공지능general purpose intelligence과는 거리가 멀어 보입니다. 이것이 두 번째 팩트입니다. 기계는 아직 특정한 상황에서 특정한 임무만을 완수하는 데 사용되고 있는데, 단지 그 임무가 조금씩 더 복잡해지고 있을 뿐입니다. 기계는 아직 '상식'과 같은 가장 기본적인 인간의 지능이 부족합니다. 인공지능은 여전히 '공포'와 같은 인간의 간단한 감정을 알지 못합니다. 두세 살 아이들도 쉽게 도와줄 수 있는 일들을 기계가 하지 못하는 경우가 많습니다.

세 번째 팩트는 이번 인공지능과 머신러닝의 응용 범위가 매우 넓다는 점입니다. 최근 인공지능과 머신러닝 기술이 곳곳에서 많이 응용되고 있는데, 이는 예전에 학술 연구 개념에 불과했던 인공지능이 이제는 대중의 시야로 들어와 미래를 결정할 수 있는 주제 중 하나로 자리매김했음을 뜻합니다. 컴퓨터 비전, 딥러닝, 로보틱스 기술, 자연어 처리 등의 기술 모두가 응용적인 측면에서 언급되고 있습니다. 여러분에게 익숙한 것들로는 안면인식, 자율 주행, 의료진단, 스마트 시티, 뉴 미디어, 게임, 교육 등이 있고, 자주 언급되진 않지만 농업 자동화, 노인 케어 시스템, 교통 통제 등에도 응용되고 있습니다. 오히려 이번 인공지능 물결의 영향을 받지 않는 사회 영역이 없을 정도입니다.

앞으로의 10년을 전망하자면, 인공지능과 머신러닝 기술은 계속해서 발전할 것이며, 관련 기술은 더욱 보편적으로 응용될 것입니다. 새로운 응용 환경도 대규모로 늘어

날 것이고, 인공지능 기초 시설 또한 빠르게 자리 잡을 것입니다. 기존의 소프트웨어나 애플리케이션은 서서히 새로운 알고리즘을 도입할 수밖에 없을 것입니다. 그렇기 때문에 저는 지금이 인공지능과 머신러닝을 배우기 가장 좋은 시기라고 생각합니다.

이 책은 어떻게 집필되었나?

국내외를 막론하고 미디어 업계는 인공지능 기술 응용의 최전선에 있습니다. 왜냐하면 미디어는 매일 천만, 심지어 억 단위의 사용자들과 만나기 때문이죠. 각양각색의 사용자들은 일상에서 콘텐츠를 떠나 살 수 없습니다. 여기서 말하는 콘텐츠에는 신문, 음악, 영화 등이 있습니다. 이런 풍부한 콘텐츠를 사용자와 연결하는 환경에는 많은 비즈니스 기회가 숨어 있습니다.

Hulu훌루는 국제적으로 가장 앞서 있는 비디오 미디어 회사입니다. 고 퀄리티의 영화와 드라마, 생방송 채널을 제공하고 있죠. Hulu 기술 아키텍처에서 가장 선진화된 부분은 인공지능과 머신러닝 알고리즘의 응용입니다. 개인화 추천, 검색, 콘텐츠 내용 이해, 콘텐츠 전송 및 재생, 광고 예측과 타기팅, 보안, 의사결정 서포트, 그리고 영상 편집과 고객 서비스까지 매우 광범위합니다. 머신러닝 알고리즘이 사용될 수 있는 배경에는 대규모 데이터 처리 시스템이 자리 잡고 있습니다. '모든 것을 알고리즘으로'는 Hulu의 현재 기술 아키텍처의 핵심이자 미래에 대한 포지셔닝 전략이기도 합니다. Hulu는 미래 IT 기술 회사이며, 모든 것을 '알고리즘화'한 앞서가는 회사이기도 합니다.

Hulu 베이징 연구소에는 각종 인공지능 알고리즘의 응용을 위해 수많은 인공지능, 머신러닝 인재들이 모여 있습니다. Hulu의 데이터 과학자, 알고리즘 엔지니어, 소프트웨어 엔지니어는 한 팀에서 일하며 매일같이 사용자의 실질적인 문제들을 해결하고 실전 경험을 쌓고 있습니다. Hulu 베이징 연구소에는 학구적인 분위기가 물씬 풍기는데, 정기적으로 머신러닝 주제를 가지고 연구 토론회를 개최하기도 하고 자체적으로 빅데이터 및 머신러닝 관련 공개 강의를 열기도 합니다.

2017년 말, 인민우전人民邮电 출판사의 위빈 편집자가 인공지능과 머신러닝 알고리즘에 대한 실용서를 쓸 수 있는지 물어 왔습니다. 현재 시장에 나와 있는 관련 서적은

크게 두 가지로 나눌 수 있습니다. 하나는 굉장히 체계적인 교과서 같은 서적이고, 다른 한 종류는 인공지능과 인류 미래에 대한 인문과학 서적입니다. 실제로 인공지능 분야에 종사하는 사람들에게 필요한 스킬을 알려주는 실용서는 많이 없었습니다.

그럼 우리가 한번 써보자는 마음으로 회사의 동료들을 대상으로 이 프로젝트에 참여할 사람을 모집했습니다. 총 15명의 책임연구원과 알고리즘 엔지니어가 이 책의 집필에 참여했고, 성공적인 프로젝트가 되었습니다. 우리는 먼저 관련 서적들을 살펴보고 브레인스토밍을 거쳐 비교적 재미있는 문답 형식의 문답집을 만들기로 했으며, 각 연구원이 관심 있어 하는 주제를 모아 기본 개념을 설명하는 데 사용했습니다.

IT 업계에서 '애자일' 개발이란 최대의 속도로 '소형화 상품'을 만들고 사용자의 피드백을 받아 상품 방향을 수정해 나가는 것을 뜻합니다. 이 책 역시 이러한 방법으로 작성되었습니다. 우리는 매주 두 개 이상의 질문을 Hulu WeChat 블로그에 게시하여 최대한 빠르게 독자들의 피드백을 얻어 출판되기 전에 많은 개선을 이뤄냈고, 단기간에 최대 효율을 달성할 수 있었습니다. 2017년 11월부터 2018년 3월까지 총 30편의 '머신러닝 문답' 시리즈를 게시했고, 이 시리즈는 업계의 호평을 받았습니다. 또한, 각종 질문과 피드백을 받게 되었는데, 이는 이 책의 핵심 내용에 포함되어 있습니다.

이 책의 장과 절 구조는 정말 많은 논의를 통해 결정된 것입니다. 인공지능, 머신러닝의 알고리즘의 범위가 너무 넓다 보니 가장 기초가 되는 내용과 개념을 포함하는 것에 초점을 맞췄고, 동시에 최신 동향까지 담으려 노력했습니다. 그렇기 때문에 이 책은 로지스틱 회귀, 의사결정 트리 등과 같은 전통적인 머신러닝 알고리즘에서부터 딥러닝, 강화학습, 앙상블 학습 등과 같은 비교적 최신 알고리즘, 그리고 학계에서 논의 중인 새로운 영역과 최신 알고리즘까지 포괄하고 있습니다. 동시에 이 책은 실전 응용 환경에서 알고리즘 시스템을 사용할 때 필요한 샘플링, 피처 엔지니어링, 모델 평가 부분을 강조하고 있습니다. 머신러닝 알고리즘은 해당 배경지식을 비교적 깊게 이해해야 하므로 각 문제와 해답 전에 해당 영역에 대한 간단한 배경 설명을 추가했습니다. 각 문제는 각기 다른 난이도를 가지고 있는데, 이는 독자들이 자신의 수준을 평가하는 데 도움이 될 것입니다.

책의 핵심이 되는 알고리즘 문답 내용 외에도 우리는 두 가지 내용을 더 추가했습니다. 하나는 '머신러닝 데이터 과학자로서의 자기 수양'으로, 업계의 전형적인 직무 내용과 요구사항을 소개하고 있습니다. 이러한 실제 예제는 많은 독자에게 인공지능 업계의 동향을 이해하고 파악하는 데 큰 도움이 될 것입니다. 두 번째는 '인공지능의 응용' 부분입니다. 많은 독자가 이미 자율 주행차, 알파고와 관련된 이야기들에 익숙할 것입니다. 우리는 업계에 종사하고 있는 사람의 시각에서 이러한 응용의 뒤에 숨겨진 원리가 무엇인지에 대해 설명하려 했습니다. 이 책을 다 읽은 후에 머신러닝 기술들을 잘 익혔다면 여러분도 이러한 현장에서 일할 수 있게 될 것입니다.

이 책은 인공지능, 머신러닝의 각 영역에 대한 많은 정보를 포함하고 있습니다. 회사에 따라, 업무에 따라, 그리고 직위에 따라 적용할 수 있는 내용도 있고, 반대로 없는 내용도 있을 것입니다. 그렇기 때문에 이 책을 읽을 때 다음과 같은 방법으로 읽을 것을 권장합니다.

1 **차례대로 읽기** 처음부터 끝까지 읽는 방법인데, 만약 모든 내용을 이해하고 모든 문제에 대해 답할 수 있다면 Hulu에 이력서를 내도 좋습니다.

2 **난이도가 낮은 것부터 읽기** 각 문제 옆에 별표로 난이도를 표시했습니다. 별 1개는 가장 간단한 문제를 뜻하고, 별 5개는 가장 어려운 문제임을 뜻합니다. 그리고 책 앞부분에서 이 책의 모든 문제와 난이도를 정리한 테이블을 제공하고 있습니다. 별 1개가 달린 문제는 기본적인 개념이나 'ROC 곡선은 무엇일까요?', '왜 수치형 특성에 대해 정규화를 진행해야 할까요?' 등과 같은 어떤 특정한 개념을 묻거나 특정 방법, 알고리즘을 사용해야 하는 이유에 관해 설명하고 있습니다. 만약 여러분이 머신러닝 입문자라면 간단한 문제에서부터 차근차근 학습해 가기를 권장합니다.

3 **목적에 따라 읽기** 각 알고리즘마다 필요한 곳이 있습니다. 모든 회사, 모든 직무에서 모든 알고리즘을 숙지할 필요는 없습니다. 만약 현재 재직 중이거나 혹은 어떤 특정한 영역에서 일하고 싶다면, 해당 직무에 필요한 알고리즘을 중점으로 공부하면 됩니다. 만약 새로운 영역에 관심이 생긴다면, 제목에서 찾아보고 해당 장을 공부하면 됩니다. 하지만 어떤 알고리즘을 사용하든 피처 엔지니어링, 모델 평가에 대한 기초 부분은 매우 중요하기 때문에 반드시 읽을 것을 권장합니다.

4 **인터넷 독서법** 한 권의 책으로 모든 내용을 깊게 다루기란 힘든 일입니다. 문제나 해답을 기반으로 확장할 수 있는 여지가 많기 때문에 각 절이나 장 마지막 부분에 '요약과 응용'을 더했습니다. 어떤 특정 분야에 관심이 있는 친구들이라면, 이 책을 시작점으로 더 깊이 있는 독서를 통해 해당 분야의 전문가가 되길 바랍니다.

5 **CEO 독서법** 만약 여러분이 관리자의 위치에 있다면, 여러분이 해결해야 하는 문제는 알고리즘이 현재 기술 시스템에 어떤 도움이 되고 어떻게 적합한 인재를 채용할 수 있을지를 판단하여 스마트화된 상품을 개발하는 것일 수도 있습니다. 그렇다면 먼저 대략적으로나마 책을 훑어보고 머신러닝 각 기술 영역에 대해 파악한 후에 적합한 해결 방안을 찾길 권합니다. 그리고 이 책을 활용해 인터뷰를 진행할 수도 있겠죠.

이 책의 출판 목적은 더 많은 사람에게 머신러닝 관련 지식을 연습하고 이해할 수 있도록 만드는 데 있습니다. 그래서 컴퓨터 관련 업계의 사람들에게 알고리즘 엔지니어링에 필요한 실제 기술을 이해시키고, 소프트웨어 개발자들이 훌륭한 데이터 과학자로 거듭날 수 있게 해 주며, 회사 관리인들에게는 인공지능 시스템에 필요한 인재와 기술을 소개하고, 인공지능, 머신러닝에 관심 있는 독자 모두에게는 기술과 시대의 최전선에 뛰어들 수 있게 도와주는 것입니다.

인공지능과 머신러닝 알고리즘은 나날이 새로워지고 있습니다. 이 책 역시 계속해서 업데이트되어야 하며 계속해서 개정판이 나와야 할 것입니다. 따라서 독자 여러분의 소중한 비판과 의견을 기다릴 것입니다. 함께 이 기술 영역의 발전을 만들어 가길 바랍니다.

베타리더 후기

제이펍은 책에 대한 애정과 기술에 대한 열정이 뜨거운 베타리더들로 하여금
출간되는 모든 서적에 사전 검증을 시행하고 있습니다.

공민서(이글루시큐리티)

이 책을 통해 미래를 대비하기 위한 멋진 가이드를 얻은 것 같습니다. 그리고 현재의 미흡한 부분도 알았습니다. 정말 고마운 책입니다. 교과서적인 전개보다는 질문과 답변 식으로 구성되어 신선했으며, 유용한 정보도 가득했습니다. 덕분에 제가 앞으로 준비해야 할 방향을 잡을 수 있었습니다.

김용현(마이크로소프트 MVP)

데이터 과학자나 개발자 채용 시 면접 및 인터뷰에 충분히 참고할 수 있는 내용이며, 머신러닝 관련 기업에 인터뷰를 준비하는 면접자와 피면접자 모두에게 도움이 되는 책입니다. 실제 AI를 깊게 활용하는 기업에서 자주 부딪히며 해결했던 문제들에 대한 기본기를 묻고 답을 제시해 주고 있어서, 입문 서적을 충분히 접한 실무자에게 색다른 방법으로 접근하는 실무 활용서로 자리매김할 것 같습니다.

노승환(크래프트테크놀로지스)

당장 인터뷰를 준비하는 분들은 물론, 처음 공부하는 분들도 흥미 있는 주제를 골라 잘 아는 분에게 물어보고 설명을 듣듯 하나하나 읽어 나가면 좋겠다고 생각했습니다. 중간중간에 나오는 '잠시 쉬어가기' 코너의 이야기들도 꿀잼이었습니다. 번역도 상당히 잘 되어 있다는 인상을 받았고, 영어 용어가 많이 병기되어 있어서 도움이 많이 될 듯합니다. 개인적으로도 예전에 공부했던 내용을 복습하는 기회가 됐습니다.

박찬성(한국전자통신연구원)

인공지능 분야의 주요 주제를 개념적으로, 그리고 수학적으로 배울 수 있는 책입니다. 책의 구성만큼이나 깔끔한 번역도 좋았습니다. 한편, 이 책은 독자가 많은 것을 알고 있다는 전제를 하고 있습니다. 따라서 내용을 빠르게 훑어보고 잘 모르는 부분에 대해서는 추가적인 자료를 검색하여 지식의 깊이를 더하면서 자신만의 지식 맵을 만든다면, 두고두고 꺼내 보며 기초를 탄탄하게 다질 수 있는 데 도움을 줄 것으로 생각합니다.

안병규

실제 인공지능 분야를 학습할 때 필요한 여러 기반 기술에 대해 친절히 설명해 놓은 책입니다. 일방적인 설명보다는 기초부터 탄탄히 설명해 나가고 있어서 이 분야를 준비하는 많은 분께 적잖은 도움이 될 것 같습니다.

양민혁(현대모비스 데이터사이언스팀)

실무에 필요한 이론, 그리고 실제 적용을 위해 고려해야 할 부분을 잘 설명하고 있습니다. 데이터 사이언스 직군의 인터뷰를 준비하는 분들에게 많은 도움이 될 것 같습니다. 저 또한 중간중간 답을 하지 못하고 막히는 부분들이 있었으나 재밌게 공부하며 읽었습니다.

이용진(삼성SDS)

이 책은 이론적인 내용과 수식을 어느 정도 알고 있는 사람들에게 적합한 내용을 담고 있습니다. '인터뷰 문답집'이라는 제목만 보면 인터뷰에서 나올 질문만 모아놓은 책이라 생각하기 쉽지만, 실제로 업무를 수행하면서 고민한 내용과 수식이 잘 정리된 책입니다. 다시 말해, 초급자보다는 중급 이상의 지식을 가진 분들에게 추천할 만한 책입니다. 전체적으로 번역도 잘 되어 있었습니다.

정태일(삼성SDS)

이 책은 인공지능과 머신러닝에 관련된 기초지식부터 실제 응용까지 폭넓게 다루고 있습니다. 인공지능과 머신러닝을 공부하다 보면 생소한 용어와 수식에 압도되거나

특정 기술을 이해하는 데 급급하여 전체적인 맥락을 보지 못하는 경우가 많은데, 이 책은 문답 형식으로 반드시 이해해야 하는 개념과 다양한 기술 간의 관계를 자연스럽게 파악할 수 있도록 합니다. 탄탄한 기본지식을 갖춘 데이터 과학자가 되는 것을 목표로 하는 분들에게 추천하고픈 책입니다. 베타리딩하며 그동안 공부했던 인공지능과 머신러닝 이론이 일목요연하게 정리되는 느낌을 받았습니다. 복잡한 수식과 어려운 용어도 있고 책 분량도 적지 않았지만, 세계적인 회사의 데이터 전문가들이 중요하다고 추려낸 문답 내용이다 보니 어느 것 하나 놓치고 싶지 않은 마음에 더 몰입하여 읽을 수 있어 좋았습니다.

조원양(하이트론씨스템즈)

이 책은 인공지능을 공부할 때 중요하지만 놓치기 쉬운 필수 지식에 대해서 인터뷰 형식으로 설명한 책입니다. 다른 기본서를 읽고 나서 이 책으로 보완하면 어느 정도 수준의 AI 엔지니어로 거듭날 수 있으리라 생각합니다. 또한, 실전에서도 사용할 수 있는 최소한의 지식을 습득할 수 있을 것입니다. 정말 흥미로운 책이었습니다.

황시연(데이터 저널리스트)

머신러닝 문제를 풀 때는 크게 두 가지로 나뉩니다. 성능이 제일 좋은 모델을 최적화하는 방법과 여러 모델의 장점을 합쳐 모델을 만드는 방법입니다. 이 책의 서술 방법은 후자에 가깝습니다. 15명의 Hulu 엔지니어가 장마다 질문에 대한 답을 실무 경험을 바탕으로 잘 녹여낸 책입니다. 질문에 대한 설명이 2페이지를 넘지 않아 이해하는 데 큰 도움이 됩니다. 그리고 책의 중간중간에 '잠시 쉬어가기' 부분은 각 주제와 관련된 핵심 개발자와 연구자들의 백그라운드에 대한 설명과 관련 지식이 나오는데, '왜' 만들었는지에 대한 생각을 할 수 있어서 유익했습니다. 또한, 각 질문은 별의 숫자로 난이도를 표시하고 있는데, 별 3개까지는 인공지능의 기초 지식이 있다면 무난하게 읽을 수 있습니다. 만약 기초지식이 부족하다면 이 책으로 인공지능 시스템이 어떻게 만들어지는지 큰 틀을 알 수 있어서 엔지니어뿐만 아니라 AI에 관심 있는 분들에게 추천해 드립니다.

CHAPTER 1
2 범주형 피처

상황 설명

범주형 피처Categorical Feature는 성별(남, 여), 혈액형(A, B, AB, O) 등과 같이 유한한 선택 범위의 값을 취하는 피처입니다. 범주형 피처는 일반적으로 문자열string 형식으로 입력되는데, 의사결정트리 등 소수의 모델이 이러한 문자열 형식 입력을 직접적으로 처리할 수 있지만, 로지스틱 회귀, 서포트 벡터 머신 등과 같은 모델에서는 반드시 수치형 피처로 전환해야 분석 작업이 가능해집니다.

키 워 드 **순서형 인코딩**Ordinal Encoding / **원-핫 인코딩**One-Hot Encoding / **이진 인코딩**Binary Encoding

질문 데이터 정제 작업을 진행할 때 범주형 피처는 어떻게 처리해야 할까요?

난이도 ★★

분석·해답

● 순서형 인코딩

순서형 인코딩ordinal encoding은 클래스 사이에 대소 관계가 존재하는 데이터에서 많이 사용됩니다. 예를 들면, 성적은 낮음, 중간, 높음으로 나눌 수 있으며, '높음 > 중간 > 낮음'과 같은 순서형 관계를 가집니다. 순서형 인코딩은 대소 관계에 따라 범주형 피처에 하나의 수치 ID를 부여합니다. 예를 들면, '좋음'은 3으로, '중간'은 2로, '낮음'은 1로 표현할 수 있습니다. 이렇게 변환된 후에도 여전히 대소 관계를 유지합니다.

● 원-핫 인코딩

원-핫 인코딩one-hot encoding은 클래스 사이에 대소 관계가 존재하지 않는 데이터에서 많이 사용됩니다. 예를 들면, 혈액형은 네 가지 값을 가질 수 있는데(A형, B형, AB형, O형), 원-핫 인코딩은 혈액형을 4차원의 희소 벡터sparse vector로 만듭니다. A형은

(1, 0, 0, 0)으로, B형은 (0, 1, 0, 0)으로, AB형은 (0, 0, 1, 0)으로, 그리고 O형은 (0, 0, 0, 1)로 표현됩니다. 클래스(유형)가 많은 경우에 원-핫 인코딩을 사용한다면 몇 가지 주의해야 할 점이 있습니다.

1 희소 벡터를 활용하여 공간을 절약해야 합니다 원-핫 인코딩을 사용하면 피처 벡터는 특정 값이 1이 되고, 기타 위치의 값들은 모두 0이 됩니다. 따라서 벡터의 희소 표현을 이용하여 공간을 절약할 수 있습니다. 또한, 현재 대부분의 알고리즘이 모두 희소 벡터 형식의 입력을 받을 수 있습니다.

2 피처 선택을 통해 최대한 차원을 줄여야 합니다 고차원 피처는 몇 가지 문제를 야기합니다. 첫 번째로, K-최근접 이웃K-Nearest Neighbor, kNN 알고리즘에서 고차원 공간하의 두 점 사이의 거리는 측정이 매우 어렵습니다. 두 번째로, 로지스틱 회귀 모형에서 파라미터의 개수가 차원에 증가에 따라 함께 증가합니다. 이는 모델의 과적합 문제를 일으킵니다. 마지막으로, 모든 차원(변수)이 분류, 예측에 도움이 되는 것은 아닙니다. 따라서 피처 선택과 함께 차원의 수를 줄여야 합니다.

이진 인코딩

이진 인코딩binary encoding은 두 단계로 나눌 수 있습니다. 첫 번째 단계는 순번 인덱스를 사용하여 각 클래스에 ID를 부여하는 것이고, 두 번째 단계는 각 클래스 ID를 이진법 코드로 나타내는 것입니다. A, B, AB, O형 혈액형을 예로 들면, 표 1.1을 통해 그 과정을 쉽게 알 수 있습니다.

표 1.1 이진 인코딩과 원-핫 인코딩

혈액형	클래스 ID	이진법 표현	원-핫 인코딩
A	1	0 0 1	1 0 0 0
B	2	0 1 0	0 1 0 0
AB	3	0 1 1	0 0 1 0
O	4	1 0 0	0 0 0 1

A형의 ID는 1이고, 이진법으로는 001로 표현됩니다. B형의 ID는 2이고, 이진법으로는 010으로 표현됩니다. 같은 방식으로 AB형과 O형의 이진법 표현도 얻을 수

있습니다. 이러한 이진법 코드는 사실상 이진법을 이용하여 ID에 대해 해시 매핑 hash mapping을 진행한 것입니다. 최종적으로는 0/1 고유벡터를 얻고, 차원의 수는 원-핫 인코딩보다 작아 저장 공간을 절약할 수 있습니다.

이번 절에서 소개한 인코딩 방법 외에 다른 인코딩 방법에는 Helmert Contrast, Sum Contrast, Polynomial Contrast, Backward Difference Contrast 등 다른 방법들도 있으니 관심이 있는 독자들은 찾아보기 바랍니다.

CHAPTER 1

3 고차원 결합 피처의 처리 방법

키워드 결합 피처Interaction Feature

질문 결합 피처란 무엇일까요? 고차원 결합 피처는 어떤 방식으로 피처 엔지니어링 해야 할까요?

난이도 ★★

분석·해답

데이터 사이에 존재하는 복잡한 관계를 보다 잘 적합fitting하기 위해 피처 엔지니어링 단계에서 일차원의 불연속 특성discrete feature을 쌍으로 조합시켜 고차원의 결합 피처interaction feature*로 만드는 작업을 진행합니다. 광고 배너에 대한 클릭 여부(선택) 예측 문제를 예로 들어, 초기 데이터에 언어와 콘텐츠 유형이라는 두 가지 이산 변수가 존재한다고 가정해 봅시다. 표 1.2는 언어와 콘텐츠 유형이 클릭 여부에 미치는 영향을 정리한 것입니다. 적합 능력을 향상시키기 위해 언어와 콘텐츠 유형을 묶어 2차 피처로 만들었고, 결과는 표 1.3에 나와 있습니다.

표 1.2 언어와 콘텐츠 유형이 클릭 여부에 미치는 영향

클릭 여부	언어	콘텐츠 유형
0	한국어	영화
1	영어	영화
1	한국어	드라마
0	영어	드라마

* 옮긴이 '상호작용 특성' 혹은 '중복 특징'이라고도 번역합니다. 여기서는 피처(특성) 사이의 결합에 초점을 두어 '결합 피처'라고 번역했습니다. 이는 두 개의 특징을 결합하여 새로운 특징을 만드는 피처 엔지니어링 방법의 한 종류입니다. 결합 피처를 만들 경우 특성(feature) 수가 늘어나기 때문에 해당 작업을 자동으로 처리할 경우 특성 폭발(feature explosion) 현상이 일어날 수 있으므로 주의해야 합니다.

CHAPTER 1

7 이미지 데이터가 부족할 때는 어떻게 처리해야 할까요?

상황 설명

머신러닝에서 대부분의 모델은 대량의 데이터를 통해 훈련되어야 합니다. 그러나 현실에서는 훈련 데이터 부족 문제가 자주 발생하게 됩니다. 예를 들어, 이미지 분류 문제는 컴퓨터 비전 영역의 가장 기초적인 문제 중 하나인데, 각 이미지를 사전에 정의한 유형 집합에 분류하는 것을 핵심 목표로 합니다. 이미지 분류 모델을 훈련할 때, 만약 훈련 데이터 샘플 수가 비교적 적다면 어떤 방법을 사용해 이 문제를 해결할 수 있을까요?

키 워 드 **전이학습**Transfer Learning / **생성적 적대 신경망**Generative Adversarial Networks, GANs / **이미지 처리**Image Processing / **업샘플링 테크닉**Up-sampling Technique / **데이터 확장**Data Augmentation

질문 이미지 분류 문제에서 훈련 데이터가 부족하다면 어떤 문제를 일으킬까요? 어떻게 데이터 부족이 야기하는 문제들을 완화할 수 있을까요?

난이도 ★★

분석·해답

모델이 제공받는 정보의 근원은 크게 두 가지가 있습니다. 첫 번째는 훈련 데이터 내에 포함된 정보이고, 두 번째는 모델 형성 과정에서 (구조, 학습, 추론 등) 사람들이 제공한 선험적 정보입니다. 훈련 데이터가 부족하다는 것은 데이터에서 얻을 수 있는 정보의 양이 제한적이라는 것을 뜻합니다. 이러한 상황에서 모델의 성능을 보장하기 위해 더 많은 선험적 정보들이 필요합니다. 선험적 정보들을 모델상에서 활용하는 방법은 특정한 내부 구조를 사용하거나 조건이나 제약 조건들을 추가하는 방법이 있습니다. 혹은 선험적 정보를 직접 데이터 세트에 포함하는 방법도 가능한데, 즉 특정한 선험적 가설에 기반해 데이터를 조정하거나 변환하거나 확장합니다. 이런 과정을 거쳐 더 유의미하고 많은 정보를 모델 훈련과 학습에 사용할 수 있습니다.

다시 문제로 돌아간다면, 이미지 분류 문제에서 훈련 데이터가 부족하다면 과적합 문제가 발생할 것입니다. 즉, 모델이 훈련 데이터상에서는 좋은 성능을 보여줄 수 있지만, 테스트 데이터상에서의 일반화 성능이 저조해지는 문제입니다. 위에서 논의한 것처럼 이 문제에 대한 해결 방안은 크게 두 가지가 있습니다. 첫 번째는 모델에 기반을 둔 방법인데, 모델의 간략화(비선형모델을 선형모델로 바꾸는 등), (L1, L2와 같은) 정규화 항 추가, 앙상블 학습 사용, 드롭아웃dropout 하이퍼파라미터 설정 등이 이 방법에 속합니다. 두 번째 방법은 데이터에 기반을 둔 방법인데, 대표적으로 데이터 확장Data Augmentation 방법이 있습니다. 즉, 선험적 지식에 기반해 특정한 정보를 남긴다는 전제하에서 초기 데이터를 변환시켜 데이터 확장 효과를 얻습니다. 이미지 분류 문제를 예로 들면, 이미지의 클래스를 변형시키지 않는 전제하에서 훈련 데이터 세트의 각 이미지에 대해 다음과 같은 변형을 진행할 수 있습니다.

❶ 일정 범위 내에서 이미지에 대해 회전, 평행 이동, 축소, 확대, 삭제, 추가, 좌우 전환 등의 변화를 줄 수 있습니다.

❷ 이미지에 대해 노이즈를 추가합니다. 예를 들면, 소금 & 후추 노이즈salt & pepper noise*, 가우스 노이즈 등이 있습니다.

❸ 색상을 변환합니다. 예를 들어, 이미지의 RGB 색상 공간상에서 주성분분석을 진행하면 세 가지 주성분의 고유벡터 p_1, p_2, p_3와 이에 대응하는 고윳값 $\lambda_1, \lambda_2, \lambda_3$을 얻습니다. 그리고 각 화소의 RBG 값에 $[p_1, p_2, p_3]\ [\alpha_1\lambda_1, \alpha_2\lambda_2, \alpha_3\lambda_3]^T$을 추가하는데, 여기서 $\alpha_1, \alpha_2, \alpha_3$는 평균이 0이고 분산이 비교적 작은 가우스 분포를 따르는 난수입니다.

❹ 이미지의 명암, 해상도, 광도, 첨예도** 등을 변환합니다.

그림 1.4는 위에서 설명한 다양한 방법에 대한 예시를 보여주고 있습니다.

---

* 옮긴이 소금 & 후추 노이즈라는 독특한 이름을 붙인 이유는 잡음(noise)이 마치 소금과 후추처럼 흰색 또는 검정색으로 이루어지기 때문입니다. 조금 더 구체적으로 설명하자면, 입력 영상의 임의의 좌표 픽셀값을 0 또는 255로 만든 형태의 잡음입니다.

** 옮긴이 acutance의 번역어로, 이미지(image area)의 명확함(sharpness)을 나타내는 척도입니다.

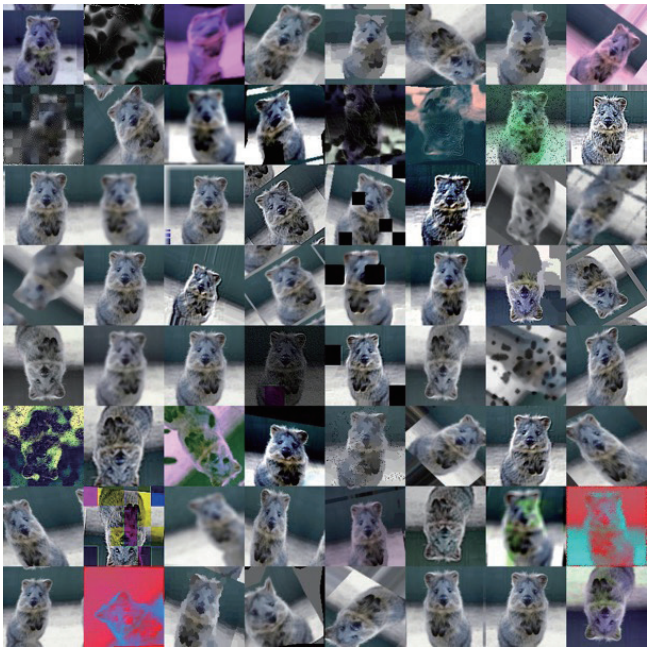


그림 1.4 이미지 데이터 확장 예시

이미지 공간에서 직접 작업하는 방식 외에 먼저 이미지의 피처를 추출하여 피처 공간 내에서 변환을 진행하고, 일반적인 데이터 확장 혹은 SMOTESynthetic Minority Over-sampling Technique와 같은 업샘플링up-sampling을 이용하는 방법도 있습니다. 이러한 휴리스틱heuristic한 변환 방법에서 벗어나고 싶다면, 오늘날 가장 유행하고 있는 생성적 적대 신경망Generative Adversarial Networks, GANs과 같은 생성모델을 활용하여 새로운 샘플을 추가하는 방법도 있습니다.

이외에도 이미 있는 다른 모델이나 데이터를 빌려와 전이학습Transfer Learning을 진행하는 방법도 딥러닝에서 자주 볼 수 있습니다. 예를 들면, 대부분의 이미지 분류 문제를 처음부터 모델링을 하지 않는 것입니다. 즉, 대규모 데이터를 통해 훈련된 기존의 좋은 모델들을 이용하여 소규모 데이터상에서 파인 튜닝fine tuning만을 진행합니다. 이러한 미세한 조정도 일종의 간단한 전이학습이라 할 수 있습니다.

CHAPTER

2

모델 평가

The Quest for Machine Learning

'측량할 수 없다면 과학이 아니다.' 이는 러시아 과학자 멘델레예프Mendeleev의 명언입니다. 컴퓨터 과학, 특히 머신러닝 영역에서 모델에 대한 평가는 매우 중요합니다. 문제와 관련 있는 평가 방법을 선택해야만 모델 선택 혹은 훈련 과정에서 일어나는 문제를 신속하게 파악하고 모델을 최적화할 수 있습니다. 모델 평가는 주로 온라인 평가와 오프라인 평가의 두 단계로 나눌 수 있습니다. 분류, 수열, 회귀, 순서예측 등 서로 다른 유형의 머신러닝 문제에 따라 평가 지표의 선택도 다를 수밖에 없습니다. 각 평가 지표의 정확한 정의를 이해하는 것, 적절한 평가 지표를 선택하는 것, 평가 지표 피드백에 기반하여 모델을 조정하는 것, 이러한 것들이 머신러닝 모델 평가 단계에서 중요한 문제들이 됩니다. 따라서 데이터 과학자라면 반드시 숙지하고 있어야 할 부분이라 할 수 있습니다.

CHAPTER 2

1 평가 지표의 한계

상황 설명

모델 평가 과정에서 분류 문제, 배열 문제, 회귀 문제 등 문제 유형에 따라 서로 다른 지표를 사용하여 평가를 진행합니다. 많은 평가 지표 중에서 대부분의 지표는 모델의 일부분 성능에 대한 단편적인 부분만을 반영합니다. 만약 평가 지표를 적절하게 사용할 수 없다면, 모델 자체의 문제를 발견할 수 없을 뿐만 아니라 잘못된 결론을 내릴 수도 있습니다. 이번 절에서는 Hulu* 팀의 실무 환경을 배경으로 몇 가지 모델 평가 환경을 가정하고, 모델 평가 지표의 한계에 대해 알아보겠습니다.

키워드 **정확도**Accuracy / **정밀도**Precision / **재현율**Recall / **평균제곱근오차**Root Mean Square Error, RMSE

질문 1 정확도의 한계성

난이도 ★

Hulu의 명품 브랜드 광고주들은 그들의 광고를 명품을 살 만한 사용자들에게만 노출하고 싶습니다. Hulu는 다른 데이터 매니지먼트 플랫폼Data Management Platform, DMP를 통해 명품 사용자들에 대한 데이터를 수집하고 훈련 세트, 테스트 세트로 나눠 명품 사용자들에 대한 분류 모델을 만들었습니다. 이 모델의 분류 정확도는 95%를 넘었지만, 실제 광고를 진행해 본 결과, 해당 모델이 광고 대부분을 명품을 구매하지 않는 사용자들에게 노출했다는 사실을 알게 됩니다. 어떤 문제 때문에 이런 상황이 발생하게 된 것일까요?

★ 옮긴이 Hulu는 미국의 OTT 서비스를 제공하는 엔터테인먼트 기업입니다. 넷플릭스(Netflix)와 유사한 서비스를 제공하는 경쟁사라고 이해하면 됩니다. 2020년 현재 약 3,000만 명의 구독자를 보유하고 있습니다.

분석·해답

이 문제에 대한 답을 알아보기 전에 먼저 분류 정확도의 정의에 대해 복습해 봅시다. 정확도accuracy는 정확하게 분류된 샘플 개수를 총 샘플 개수로 나눈 것입니다. 즉, 다음의 식과 같습니다.

$$Accuracy = \frac{n_{\text{correct}}}{n_{\text{total}}} \tag{2.1}$$

여기서 n_{correct}는 정확하게 분류된 샘플 개수이고, n_{total}은 전체 샘플 개수입니다.

정확도는 분류 문제에서 가장 간단하고 직관적인 평가 지표이지만 명백한 단점이 있습니다. 예를 들어, 샘플의 구성이 잘못되어 답이 아닌 음성negative 샘플이 전체 데이터에서 99%를 차지하고 있다면, 분류기가 모든 샘플에 대한 예측값을 음성으로 예측할지라도 99%의 정확도를 나타내게 됩니다. 그렇기 때문에 클래스별로 샘플 비율이 불균형한 경우, 정확도는 불균형 데이터의 영향을 많이 받게 됩니다.

이 부분을 이해했다면 앞서 말한 문제에 대한 해답을 쉽게 찾을 수 있을 것입니다. 명품 구매 사용자가 Hulu 전체 사용자 수에서 차지하는 비중이 매우 적기 때문에 모델의 전체 분류 정확도가 높다 할지라도 명품 구매 사용자에 대한 분류 정확도가 높다고는 할 수 없습니다. 플랫폼에서 광고를 노출할 때 우리는 모델이 사전에 판별한 '명품 구매 사용자'들에게만 광고를 노출하게 되는데, '명품 구매 사용자'들에 대한 판별 정확도가 높지 않기 때문에 이런 현상이 발생한 것입니다. 이 문제를 해결하기 위해서는 평균 정확도(각 클래스 샘플의 정확도의 산술평균)를 모델 평가 지표로 사용할 수 있습니다.

사실, 이런 문제는 개방형 면접 질문(정답이 여러 개가 존재하는 문제)입니다. 면접자는 당면한 문제에 대해 차근차근 원인을 찾아가면 됩니다. 문제에 대한 정답은 지표 선정 오류만이 아닙니다. 예를 들어, 평가 지표를 맞게 선택했다 하더라도 과적합이나 과소적합이 존재할 수도 있고, 테스트 세트와 훈련 세트를 제대로 분류하지 않았기 때문에 발생하는 문제일 수도 있습니다. 또한, 오프라인 평가와 온라인 평가 샘플 분포에 차이가 존재하기 때문일 수도 있습니다. 그러나 평가 지표의 선택은 가장 쉽게 발견할 수 있는 원인이고, 평가 결과에 가장 큰 영향을 미칠 수 있는 요소입니다.

질문 2 정밀도와 재현율의 균형

난이도 ★

Hulu는 콘텐츠 연관 검색 기능을 제공합니다. 연관 검색 랭킹 모델이 출력하는 TOP 5 콘텐츠의 정밀도는 매우 높습니다. 그러나 실제 사용 과정에서 사용자들은 찾고 싶어 하는 영상을 못 찾는 경우가 많다고 합니다. 특히, 비교적 인기가 없는 콘텐츠의 경우에 더욱더 그렇습니다. 이 문제의 원인은 무엇일까요?

분석·해답

이 문제에 답하기 위해서는 먼저 정밀도와 재현율, 이 두 가지 개념을 명확히 해야 합니다. 정밀도precision는 분류기가 양성 샘플이라고 분류한 것 중에서 실제 양성 샘플인 것의 비율입니다. 재현율recall은 실제 양성 샘플인 것 중에서 분류기가 정확히 분류해 낸 양성 샘플의 비율입니다.

일반적인 랭킹 문제에서는 얻은 결과에 대해 직접적으로 양성 샘플 혹은 음성 샘플을 판별하는 정해진 임곗값이 없습니다. 대신, Top N으로 반환된 결과의 정밀도 값과 재현율 값으로 랭킹 모델의 성능을 평가합니다. 문제에서 설명한 것처럼 사용자가 찾고 싶은 콘텐츠를 찾지 못하는 현상이 잦다면, 이는 모델이 관련성 있는 콘텐츠를 충분히 찾아 주지 못했다는 것을 뜻합니다. 이것은 재현율이 낮다는 뜻입니다. 만약 관련 결과가 100개라고 가정한다면, Precision@5가 100%일 때 Recall@5는 5%가 됩니다. 모델을 평가할 때 정밀도와 재현율을 동시에 고려해야 할까요? 다시 말해, 서로 다른 Top N의 결과들에 대해 관찰해야 할까요? 아니면 더 고차원적인 평가 지표로 정밀도와 재현율을 골고루 고려해야 할까요?

위에서 한 질문들은 모두 정답입니다. 종합적으로 랭킹 모델을 평가하기 위해서는 서로 다른 Top N하에서의 Precision@N과 Recall@N을 고려해야 하는데, 가장 좋은 방법은 P-RPrecision-Recall 곡선을 그려 보는 것입니다. 따라서 P-R 곡선을 그리는 방법에 대해 간단히 설명하고 넘어가겠습니다.

P-R 곡선의 x축은 재현율, y축은 정밀도입니다. 랭킹 모델의 예에서, P-R 곡선상의 하나의 점은 어떠한 임곗값에서 모델이 해당 임곗값보다 큰 결과는 양성 샘플로 판별하고 해당 임곗값보다 작은 결과는 음성 샘플로 판단하는데, 이때 반환된 결과에

대응하는 재현율과 정밀도를 뜻합니다. P-R 곡선은 임곗값을 높은 곳에서 낮은 곳으로 이동시키며 만들어집니다. 그림 2.1은 P-R 곡선의 샘플 그래프인데, 여기서 실선은 모델 *A*의 P-R 곡선을 나타내고, 점선은 모델 *B*의 P-R 곡선을 나타냅니다. 원점 주변은 임곗값이 가장 클 때의 모델의 정밀도와 재현율을 나타냅니다.

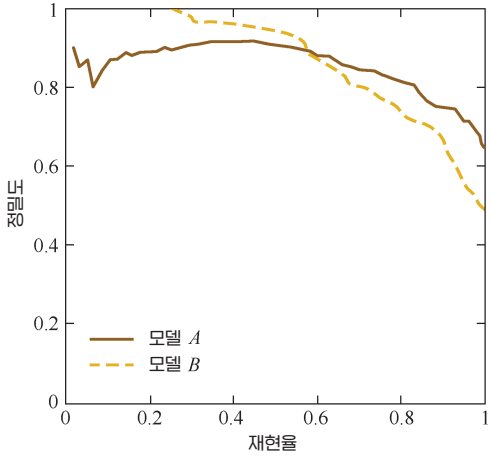


그림 2.1 P-R 곡선의 샘플 그래프

그림을 통해 알 수 있는 점은 재현율이 0에 가까울 때 모델 *A*의 정밀도는 0.9이고 모델 *B*의 정밀도는 1입니다. 이는 모델 *B*에서 순위에 랭크된 샘플들은 모두 실제 양성 샘플이라는 것을 뜻하고, 반대로 모델 *A*의 경우에는 고득점을 얻은 몇 개의 샘플에 잘못 예측할 가능성이 존재한다는 것을 뜻합니다. 그리고 재현율이 증가함에 따라 정밀도는 전체적으로 내려갑니다. 그러나 재현율이 1일 때 모델 *A*의 정밀도는 모델 *B*의 정밀도를 능가합니다. 이는 어떤 점 위에 대응하는 정밀도와 재현율만 고려해서는 모델의 성능을 완벽하게 측정하기 힘들다는 사실을 설명해 줍니다. P-R 곡선의 전체적인 표현을 확인해야 모델에 대한 전면적인 평가가 가능할 것입니다.

이 외에 F1 score와 ROC 곡선도 랭킹 모델의 성능을 종합적으로 반영할 수 있습니다. F1 score는 정밀도와 재현율의 조화 평균이고, 다음과 같이 정의됩니다.

$$\mathrm{F1} = \frac{2 \times precision \times recall}{precision + recall} \tag{2.2}$$

ROC 곡선에 대해서는 별도의 절에서 다루기 때문에 여기서는 설명하지 않겠습니다.

질문 3 평균제곱근오차의 예외

난이도 ★

Hulu는 스트리밍 서비스 회사이기 때문에 다양한 미국 드라마를 보유하고 있습니다. 각 드라마의 시청률 추세를 예측하여 타깃 광고를 진행하는 것은 매우 중요한 부분입니다. 따라서 우리는 어떠한 드라마의 시청률 추세를 예측하는 회귀모델을 만들고 싶습니다. 그러나 어떤 종류의 회귀모델을 쓰든지 얻게 되는 RMSERoot Mean Square Error 지표가 모두 매우 높게 나타납니다. 그런데 95%의 시간대 내에서 모델의 예측오차는 1%가 채 되지 않습니다. 예측오차만 살펴보면 매우 좋은 예측 결과라고 할 수 있는데, 오차 대비 너무 높은 RMSE 지표가 나타나는 가장 가능성 높은 원인은 무엇일까요?

분석·해답

RMSE는 회귀모델을 평가할 때 자주 사용되는 지표입니다. 하지만 위에서 서술한 상황을 볼 때 RMSE 지표는 측정 효과를 상실한 것 같습니다. 먼저 RMSE 계산 공식에 대해 살펴보겠습니다.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{2.3}$$

여기서 y_i는 i번째 샘플의 실제 값이고, $\hat{y}_i$는 i번째 샘플의 예측값입니다. n은 샘플의 개수입니다.

일반적인 상황에서 RMSE는 회귀모델의 예측값이 실제 값에서 벗어난 정도를 잘 반영합니다. 하지만 벗어난 정도가 매우 큰 특이점outlier, 이상점이 존재할 경우에는 이 소수의 특이점 때문에 RMSE 지표가 매우 높아지게 됩니다.

문제에서 모델은 95% 시간대 내에서 예측오차가 1% 미만이라고 했습니다. 이는 대부분의 시간대 내에서 모델의 예측 성능이 매우 뛰어남을 뜻합니다. 그러나 RMSE 지표는 매우 저조한데, 이는 기타 5% 시간대 내에 심각한 특이점이 존재할 가능성이 높다는 것을 뜻합니다. 사실, 유입량이나 시청률 예측 문제에서는 노이즈noise 포인트가 매우 쉽게 발견됩니다. 예를 들어, 시청률이 매우 저조한 드라마도 있을 것이고, 막 시작을 했거나 최근에 상을 받은 드라마도 있을 것입니다. 이러한 갑작스러운 이벤트로 인해 유입된 사용자들이 특이점 데이터 그룹을 형성할 가능성이 높습니다.

그렇다면 이 문제에 대한 해결 방안은 무엇이 있을까요? 세 가지 각도로 접근이 가능한데, 먼저 이러한 특이점들이 단순 노이즈라면 데이터 전처리 과정에서 필터링하는 방법이 있습니다. 두 번째로 특이점들이 단순 노이즈가 아니라고 판단된다면 모델의 예측 성능을 향상시켜 특이점 데이터들이 만들어 낸 메커니즘을 모델에 포함시켜야 합니다(이 토픽은 매우 범위가 넓기 때문에 여기서 다 설명하지는 않겠습니다). 마지막으로, 더 적절한 평가 지표를 사용하는 방법이 있습니다. 평가 지표에 대해서 얘기하면, 사실 RMSE보다 더 견고한robust 지표가 있는데, 예를 들면 평균절대비오차Mean Absolute Percent Error, MAPE가 있습니다. MAPE는 다음과 같이 정의됩니다.

$$MAPE = \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times \frac{100}{n} \tag{2.4}$$

RMSE와 비교했을 때 MAPE는 각 포인트의 오차들을 정규화하여 각 특이점에 의해 발생하는 절대 오차의 영향을 낮출 수 있습니다.

요약·응용

이번 절에서는 Hulu 플랫폼 운영 환경에서 나올 수 있는 문제와 상황을 가정하여 적절한 평가 지표의 중요성에 관해 설명했습니다. 평가 지표마다 고유의 가치가 존재합니다. 하지만 단일 평가 지표만을 사용하여 모델을 평가할 경우, 단편적인 결론을 내릴 뿐만 아니라 심지어 잘못된 결론을 내릴 수도 있습니다. 따라서 서로 보완성이 존재하는 지표들을 종합적으로 고려하여 모델을 평가해야 더 쉽고 정확하게 모델에 존재하는 문제들을 찾아낼 수 있습니다. 실제 비즈니스 문제를 풀 때 매우 중요한 부분이기 때문에 잘 숙지할 수 있도록 합시다.

CHAPTER 2

ROC 곡선

상황 설명

이진분류기Binary Classifier는 머신러닝 영역에서 가장 자주 보이고 광범위하게 응용되는 분류기입니다. 이진분류기를 평가하는 지표는 정밀도, 재현율, F1 score, P-R 곡선 등이 있습니다. 이전 절에서 이미 이러한 지표에 대해 간단한 소개를 했었는데, 정도에 따라 다르지만 소개했던 지표들이 모델의 일부 성능만 반영한다는 한계점도 느끼셨을 겁니다. 이에 비해 ROC 곡선은 많은 장점을 가지고 있습니다. 따라서 이진분류기를 평가하는 가장 중요한 지표 중 하나라고 할 수 있습니다. 그럼, 다음 질문들을 통해 ROC 곡선을 그리는 방법과 그 특징에 대해 살펴보겠습니다.

키워드 ROC 곡선Receiver Operating Characteristic Curve / AUCArea Under Curve / P-R 곡선Precision-Recall Curve

질문 1 ROC 곡선이란 무엇일까요?

난이도 ★

분석·해답

ROC 곡선은 Receiver Operating Characteristic Curve의 약자입니다. ROC 곡선은 원래 군사 영역에서 유래된 개념으로, 나중에는 의학 영역에서 발전하였습니다. '수신자 조작 특성 곡선'이라는 명칭도 의학 영역에서 유래된 것입니다.

ROC 곡선의 가로축은 거짓 양성 비율False Positive Rate, FPR을 나타내고, 세로축은 실제 양성 비율True Positive Rate, TPR을 나타냅니다. FPR과 TPR의 계산 방법은 각각 다음과 같습니다.

$$FPR = \frac{FP}{N} \tag{2.5}$$

$$TPR = \frac{TP}{P} \tag{2.6}$$

위 식에서 P는 실제 양성 샘플 수를 뜻하고, N은 실제 음성 샘플 수를 뜻합니다. TP는 P개의 양성 샘플 중에서 분류기가 양성 샘플로 예측한 샘플의 개수를 나타내고, FP는 N개의 음성 샘플 중에서 분류기가 양성 샘플로 예측한 샘플의 개수를 나타냅니다.

정의만 봐서는 조금 헷갈리기 쉽습니다. 더 직관적으로 설명하기 위해 많이 사용하는 환자 진단 예를 통해 다시 설명하겠습니다. 먼저, 10명의 암 의심 환자가 있는데 여기서 3명만이 실제 암에 걸렸다고 가정해 봅시다(P = 3). 그 외 7명은 암에 걸리지 않았습니다(N = 7). 병원에서 10명의 환자에 대한 진단을 해서 3명의 암환자가 있다고 결론을 내렸습니다. 하지만 여기서 실제 암환자는 2명뿐입니다(TP = 2). 그렇다면 실제 양성 비율 $TPR = TP/P = 2/3$을 계산할 수 있습니다. 불행하게도, 7명의 암에 걸리지 않은 환자들 중 한 명이 오진을 받았습니다(FP = 1). 그렇다면 거짓 양성 비율 $FPR = FP/N = 1/7$을 계산할 수 있습니다. 해당 병원의 진단 자체를 하나의 분류기로 생각한다면, 이 분류기의 분류 결과는 ROC 곡선상의 점 (1/7, 2/3)이 됩니다.

질문 2 ROC 곡선은 어떻게 그릴까요?

난이도 ★★

분석·해답

ROC 곡선은 분류기의 '절단점'을 계속해서 이동하며 곡선상의 중요 지점을 생성합니다. 다음 예제를 통해 '절단점' 개념에 대해 알아보겠습니다.

이진분류 문제에서 모델의 출력은 일반적으로 샘플이 양성일 확률입니다. 예를 들어, 테스트 세트에 20개의 샘플이 있고 표 2.1과 같은 결과를 출력했다고 가정해 봅시다. 샘플은 예측확률이 높은 순서대로 정렬되었습니다. 모델을 양성, 음성의 값으로 출력하기 전에 임곗값을 정해 주어야 합니다. 예측확률이 임곗값보다 높다면 양성으로 판별되고, 임곗값보다 작다면 음성으로 분류됩니다. 예를 들어,

지정 임곗값이 0.9라면 첫 번째 샘플만이 양성으로 예측되고, 나머지는 모두 음성으로 예측될 것입니다. 앞서 말한 '절단점*'이란 바로 양성과 음성 예측 결과를 구별하는 임곗값을 뜻합니다.

절단점은 동적으로 조절할 수 있는데, 높은 점수부터 시작해서 낮은 점수로 이동시키고, 각 절단점은 모두 하나의 FPR과 TPR에 대응합니다. ROC 그림에서 각 절단점에 대응하는 위치를 그리고 모든 점을 연결하면 최종적으로 ROC 곡선을 얻을 수 있습니다.

표 2.1 이진분류 모델의 출력 결과 샘플

샘플 인덱스	실제 레이블	모델 출력확률	샘플 인덱스	실제 레이블	모델 출력확률
1	p	0.9	11	p	0.4
2	p	0.8	12	n	0.39
3	n	0.7	13	p	0.38
4	p	0.6	14	n	0.37
5	p	0.55	15	n	0.36
6	p	0.54	16	n	0.35
7	n	0.53	17	p	0.34
8	n	0.52	18	n	0.33
9	p	0.51	19	p	0.30
10	n	0.505	20	n	0.1

이번 예제에서 절단점이 무한대일 경우 모델은 모든 샘플을 음성으로 예측합니다. 그렇게 된다면 FP와 TP는 모두 0이 됩니다. 그리고 FPR과 TPR도 모두 0이 됩니다. 따라서 곡선의 첫 번째 점의 좌표는 (0, 0)이 됩니다. 절단점을 0.9로 조절하면 모델은 1번 샘플을 양성 샘플로 예측하게 됩니다. 이 샘플은 실제로도 양성 샘플입니다. 따라서 TP는 1이 되고, 20개의 샘플 중에서 양성 샘플의 수는 10(P = 10)이기 때문에 $TPR = TP/P = 1/10$을 얻습니다. 여기서 잘못 예측한 양성 샘플이 없기 때문

* 옮긴이 임곗값(threshold) 혹은 컷오프(cut-off)라고 많이 부르지만, 원문의 의미를 살리기 위해 '절단점'이라고 번역했습니다.

에 FP는 0이 되고, 모든 음성 샘플 수는 10입니다($N = 10$). 따라서 $FPR = FP/N = 0/10 = 0$이 되고, ROC 곡선상의 점 (0, 0.1)이 됩니다. 이런 식으로 절단점을 조절해 가면 모든 주요 지점에 대해 나타낼 수 있고, 이들을 모두 잇는다면 그림 2.2와 같은 ROC 곡선을 얻을 수 있게 됩니다.

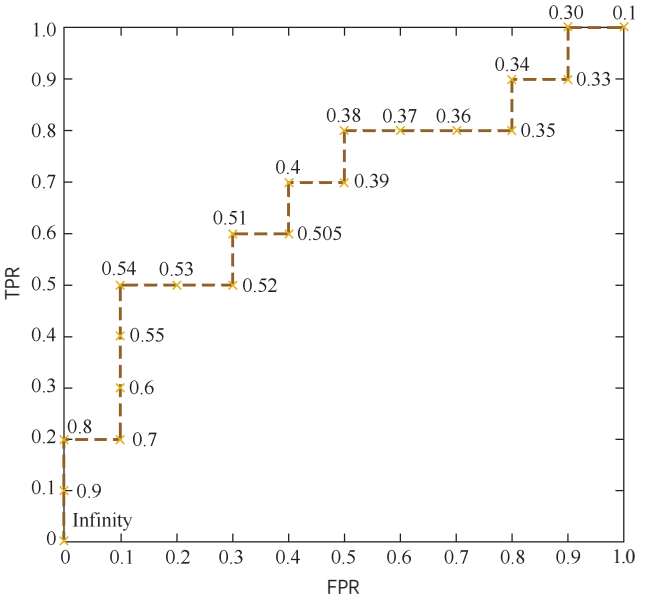


그림 2.2 ROC 곡선

사실, ROC 곡선을 그리는 더 직관적인 방법이 존재합니다. 먼저, 샘플 레이블에 기반해 양성, 음성 샘플 수를 계산합니다. 일반적으로는 양성 샘플 수를 P, 음성 샘플 수를 N으로 설정합니다. 그리고 x축의 간격을 $1/N$로 설정하고, y축의 간격을 $1/P$로 설정합니다. 그런 다음, 모델이 출력한 예측확률에 기반해 샘플을 정렬합니다(높은 순서대로). 모든 샘플을 대상으로 0에서 시작하여 ROC 곡선을 그리기 시작하면 되는데, 양성 샘플을 만날 때마다 y축 방향에서 설정한 간격에 따라 곡선 그래프를 그리고, 음성 샘플을 만날 때마다 수평축 방향에서 설정한 간격에 따라 곡선 그래프를 그립니다. 이러한 과정을 모든 샘플에 대해 진행하고 (1, 1) 점에서 멈추면 ROC 곡선을 완성할 수 있습니다.

질문 3 AUC는 어떻게 계산할까요?

난이도 ★★

분석·해답

이름에서 알 수 있듯이, AUC$_{\text{Area Under Curve}}$는 ROC 곡선 아래의 면적을 뜻합니다. 이 지표는 ROC 곡선에 기반해 모델 성능을 정량화하여 나타낼 수 있습니다. AUC 값을 계산하기 위해서는 ROC 곡선의 x축을 따라 적분만 해주면 됩니다. 대부분의 ROC 곡선은 $y = x$ 선보다 높은 곳에 위치하기 때문에 AUC의 값은 일반적으로 0.5~1 사이에 있습니다(만약 아니라면 모델이 예측한 확률을 뒤집으면$(1 - p)$ 더 좋은 분류기를 얻을 수 있게 됩니다). AUC가 클수록 분류기의 성능이 더 좋다는 것을 나타냅니다.

질문 4 ROC 곡선과 P-R 곡선을 비교해 보세요.

난이도 ★★★

분석·해답

2장 1절에서 P-R 곡선에 대해서 설명했는데, P-R 곡선과 비교했을 때 ROC 곡선은 다음과 같은 특징이 있습니다. 양성, 음성 샘플의 분포에 변화가 생겼을 때 ROC 곡선의 형태는 기본적으로 변하지 않고 유지되지만, P-R 곡선의 형태는 일반적으로 급격한 변화를 보입니다.

예를 들어, 그림 2.3은 ROC 곡선과 P-R 곡선의 비교 그래프인데, 그림 2.3(a)와 그림 2.3(c)는 ROC 곡선이고, 그림 2.3(b)와 그림 2.3(d)는 P-R 곡선입니다. 그림 2.3(c)와 그림 2.3(d)는 테스트 세트의 음성 샘플 수를 10배로 늘린 후 그린 곡선 그래프입니다.

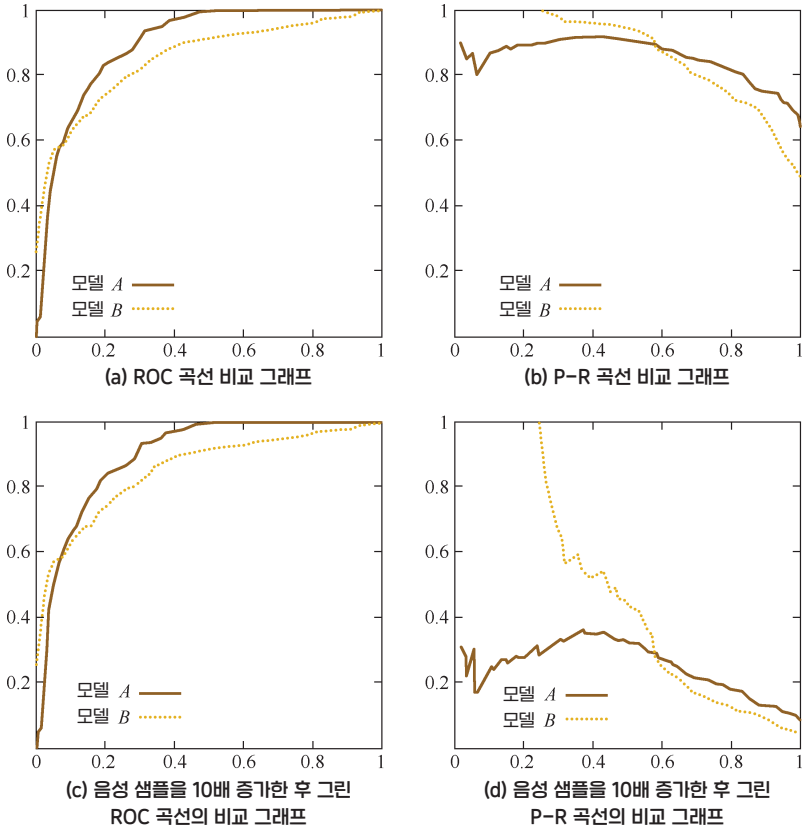


그림 2.3 ROC 곡선과 P-R 곡선 비교

위에서 확인할 수 있듯이, P-R 곡선에는 뚜렷한 변화가 확인되지만 ROC 곡선의 형태는 기본적으로 변하지 않습니다. 이러한 특징은 ROC 곡선이 다양한 테스트 세트를 만날 때마다 견고한 결과를 보여줄 수 있게 만들어 더 객관적으로 모델 자체의 성능을 평가할 수 있게 해줍니다. 이러한 특징의 실질적 의미는 무엇일까요? 많은 현실 문제에서 양성, 음성 샘플의 수는 불균형합니다. 예를 들어, 광고 영역에서 자주 사용되는 전환율 예측모델에서 양성 샘플의 수는 음성 샘플 수의 1/1,000, 심지어 1/10,000인 경우도 많습니다. 만약 다른 테스트 세트를 선택한다면 P-R 곡선의 변화는 매우 클 것이지만, ROC 곡선은 안정적으로 모델 자체의 성능을 반영할 수 있을 것입니다. 따라서 ROC 곡선은 랭킹, 추천, 광고 등 분야에서 더 자주 사용

됩니다. 하지만 주의해야 할 것은 P-R 곡선을 선택하느냐, ROC 곡선을 선택하느냐에 대한 문제는 해결하고자 하는 문제에 따라 달라진다는 것입니다. 만약 모델이 특정 데이터 세트상에서 어떤 성능을 내는지 알고 싶다면, P-R 곡선을 선택하는 것이 더 좋을 수도 있습니다.

잠시 쉬어가기...

ROC 곡선의 유래

ROC 곡선이 최초로 사용된 곳은 군사 영역입니다. 그 후 의학 영역에서 많이 사용되었고, 1980년대 후반부터 머신러닝 영역에서 사용되었습니다. 세계 2차대전 기간에 레이더병의 주요 임무는 레이더 모니터를 죽어라 쳐다보며 적군이 오는지를 확인하는 것이었습니다. 이론상으로는 적군 전투기의 기습이 있다면 레이더 모니터에 상응하는 신호가 나타나야 합니다. 그러나 실제로는 적군 전투기뿐만 아니라 새가 레이더 범위에 들어왔을 때도 신호가 발생했습니다. 만약 너무 조심스러워 신호가 나타날 때마다 적의 전투기라고 판단해 버리면 오보 위험이 커지고, 반대로 너무 관대하여 모든 신호를 새들 때문이라고 생각한다면 중요한 정보를 놓치는 위험에 처하게 되었습니다. 레이더병들은 새의 신호와 전투기 신호 사이의 차이를 구별하려 애썼는데, 문제는 레이더병마다 자기 자신만의 판단 기준이 있어서 (어떤 병사는 너무 신중하고, 어떤 병사는 너무 대담해서) 통일된 신호를 주지 못한다는 것이었습니다.

각 레이더병의 보고 정확성을 연구하기 위해 관리자는 모든 레이더병의 보고 특징을 종합했습니다. 특히, 그들이 잘못 보고하거나 누락시킨 보고에 대한 각각의 확률을 2차원 좌표계에 그렸습니다. 이 2차원 좌표의 y축은 민감성(실제 양성율)이었는데, 즉 모든 적군의 기습 사건 중에서 각 레이더병이 정확하게 예측한 확률을 나타냈습니다. 그리고 x축은 1-특이성(거짓 양성률)이었고, 모든 적군이 아닌 신호 중에서 레이더병이 잘못 보고한 확률을 나타냈습니다. 각 레이더병의 보고 기준이 다르기 때문에 얻은 민감성과 특이성의 조합도 서로 달랐습니다. 레이더병의 보고 성능에 대해 종합한 후, 관리자는 그들이 하나의 곡선상에 놓여 있다는 것을 발견했습니다. 이 곡선이 바로 의학계와 머신러닝 영역에서 자주 사용되는 ROC 곡선입니다.

CHAPTER 2
3 코사인 거리의 응용

상황 설명

이번 장의 주제는 모델 평가인데, 모델 훈련 과정에서 우리는 샘플 사이의 거리에 대해 비교하는 경우가 많습니다. '샘플 사이의 거리를 어떻게 측정할 것인가' 역시 최적화 목표와 훈련 방법의 기초가 됩니다.

머신러닝 문제에서 특성은 벡터의 형태로 표현되는 경우가 많습니다. 따라서 두 특성 벡터 사이의 유사도를 분석할 때 코사인 유사도를 자주 사용합니다. 코사인 유사도 값의 범위는 [-1, 1]이고, 같은 두 벡터 사이의 유사도는 1입니다. 만약 거리와 유사한 형태로 표현하고 싶다면 1에서 코사인 유사도를 뺀 것이 코사인 거리가 됩니다. 따라서 코사인 거리가 취할 수 있는 값의 범위는 [0, 2]가 되고, 동일한 두 벡터의 코사인 거리는 0이 됩니다.

키워드 **코사인 유사도**Cosine Similarity / **코사인 거리**Cosine Distance / **유클리드 거리**Euclidean Distance / **거리의 정의**Definition of Distance

질문 1 어떤 상황에서 유클리드 거리 대신 코사인 유사도를 사용하는지를 학습과 연구 경험을 토대로 설명해 보세요.

난이도 ★★

분석·해답

두 벡터 A와 B에 대해 코사인 유사도는 $\cos(A,B)=\frac{A\bullet B}{\|A\|_2\|B\|_2}$으로 정의됩니다.

즉, 두 벡터 사이의 코사인인데, 초점은 두 벡터 사이의 각도에 맞춰져 있지 그들의 절대 크기가 아닙니다(값의 범위는 [-1, 1]입니다). 한 쌍의 유사한 텍스트에서 길이는 많이 다르지만 내용이 비슷한 경우, 이때 단어 빈도나 단어 벡터를 특성으로 사용한다면 특징 공간에서의 유클리드 거리는 일반적으로 매우 커집니다. 그러나 코사인 유사도를 사용한다면 그들 사이의 각도가 작기 때문에 유사도가 높게 나옵니다. 이 외에도 텍스트, 이미지, 비디오 등에 데이터를 사용하는 영역들은 특성의 차원이

매우 높은 경우가 많은데, 코사인 유사도cosine similarity는 고차원 데이터에 대해서도 '방향이 같을 경우는 1, 90도의 각을 이룬다면 0, 반대 방향을 가진다면 −1'의 값을 가지지만, 유클리드 거리의 수치는 차원의 영향을 받아 값의 범위가 불안정하고 함의(담긴 뜻) 역시 비교적 모호해집니다.

벡터의 길이가 정규화된 Word2Vec에서는 유클리드 거리와 코사인 거리가 단조 관계*를 보입니다. 즉, 다음과 같습니다.

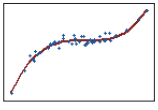
$$\| A - B \|_2 = \sqrt{2(1 - \cos(A, B))} \tag{2.7}$$

여기서 $\| A - B \|_2$는 유클리드 거리를 나타내고, $\cos(A, B)$는 코사인 유사도를 나타냅니다. 그리고 $(1 - \cos(A, B))$는 코사인 거리를 나타냅니다. 이러한 상황에서 거리가 가장 작은(유사도가 가장 높은) 이웃을 선택한다면, 코사인 유사도와 유클리드 거리 둘 중 어느 것을 사용하든 같은 결과가 나올 것입니다.

다시 정리하면, 유클리드 거리는 수치상의 절대 차이를 나타내고, 코사인 거리는 방향과 크기의 상대적 차이를 반영합니다. 예를 들어, 두 드라마 콘텐츠 사용자들의 행동 데이터를 관찰할 때, 사용자 A의 시청 벡터는 (0, 1), 사용자 B의 시청 벡터는 (1, 0)이라고 가정해 봅시다. 이때 두 사용자의 코사인 거리는 매우 길지만, 유클리드 거리는 짧습니다. 서로 다른 콘텐츠에 대한 두 사용자의 선호도를 분석하고 싶다면 우리는 상대적 차이에 더 관심을 가져야 하고, 이때는 당연히 코사인 거리를 사용해야 합니다. 그러나 우리의 목적이 사용자의 활성도일 때, 사용자의 로그인 횟수와 평균 시청 시간을 특성으로 하는 코사인 거리가 (1, 10), (10, 100)이라고 가정한다면, 두 사용자는 매우 가깝다는 결론을 내리게 됩니다. 하지만 두 사용자의 활성도는 매우 다른데, 이때 우리의 목적은 절대적 차이를 확인하는 것이므로 유클리드 거리를 사용해야 합니다.

특정 측정 방법을 어떤 문제에서 사용해야 하는가에 대한 부분은 평소 연구와 공를 하면서 계속해서 정리하고 생각해야 할 문제입니다. 그래야만 면접 기회가 주어

★ 옮긴이 단조 관계에서 변수(변량)는 동일한 방향으로 이동하지만 같은 속도로 이동하는 것은 아닙니다.



졌을 때 논리정연하게 답할 수 있으며, 새로운 문제를 만났을 때 문제 해결 능력을 갖추게 됩니다.

질문 2 코사인 거리는 엄격한 의미에서의 거리가 맞습니까?

난이도 ★★★

분석·해답

이 문제는 지원자의 거리의 정의에 대한 이해와 간단한 추론 능력을 알아보기 위한 것입니다. 먼저, 거리에 대한 정의를 알아봅시다. 하나의 집합에서 만약 각 쌍의 원소가 모두 하나의 실수로 정의될 수 있고 세 가지 거리 공식(구분 불가능한 점의 동일성positive definiteness, 대칭성, 삼각부등식)이 성립한다면, 해당 실수는 이 원소 사이의 거리라고 정의할 수 있습니다.

코사인 거리cosine distance는 동일성과 대칭성은 만족하지만, 삼각부등식을 만족하진 못합니다. 따라서 엄격한 정의에서의 거리라고 할 수 없습니다. 구체적으로 설명하면, 벡터 A와 B에 대해 세 가지 거리 공식은 다음과 같은 증명 과정을 거치게 됩니다.

- **동일성**

  코사인 거리의 정의에 의해 다음과 같은 식을 얻습니다.

$$\mathrm{dist}(A,B)=1-\cos\theta=\frac{\|A\|_2\|B\|_2-AB}{\|A\|_2\|B\|_2} \tag{2.8}$$

  $\|A\|_2\|B\|_2-AB\geqslant 0$을 고려했을 때 $\mathrm{dist}(A,B)\geqslant 0$이 항상 성립합니다. 특히,

$$\mathrm{dist}(A,B)=0\Leftrightarrow\|A\|_2\|B\|_2=AB\Leftrightarrow A=B \tag{2.9}$$

  위와 같은 성질이 있습니다. 따라서 코사인 거리는 동일성을 만족합니다.

- **대칭성**

  코사인 거리의 정의에 의해 다음과 같은 식을 얻습니다.

$$\begin{aligned}\mathrm{dist}(A,B)&=\frac{\|A\|_2\|B\|_2-AB}{\|A\|_2\|B\|_2}=\frac{\|B\|_2\|A\|_2-AB}{\|B\|_2\|A\|_2}\\&=\mathrm{dist}(B,A)\end{aligned} \tag{2.10}$$

따라서 코사인 거리는 대칭성을 만족합니다.

- **삼각부등식**

  이 성질은 성립하지 않는데, 다음에서는 하나의 반대되는 예를 보여주고 있습니다. $A = (1, 0)$, $B = (1, 1)$, $C = (0, 1)$을 가정한다면,

$$\mathrm{dist}(A,B) = 1 - \frac{\sqrt{2}}{2} \tag{2.11}$$

$$\mathrm{dist}(B,C) = 1 - \frac{\sqrt{2}}{2} \tag{2.12}$$

$$\mathrm{dist}(A,C) = 1 \tag{2.13}$$

  위 식과 같이 됩니다. 따라서 다음의 식을 얻을 수 있습니다.

$$\mathrm{dist}(A,B) + \mathrm{dist}(B,C) = 2 - \sqrt{2} < 1 = \mathrm{dist}(A,C) \tag{2.14}$$

만약 인터뷰할 때 긴장하여 이러한 계산이 머릿속에 쉽게 떠오르지 않을 때는 어떻게 해야 할까요? 이때는 코사인 거리와 유클리드 거리의 관계를 생각해 보면 됩니다. 문제 1에서 우리는 단위원unit circle(반지름의 길이가 1인 원을 뜻함)에서 유클리드 거리와 코사인 거리가 다음을 만족한다는 것을 알고 있습니다.

$$\| A - B \| = \sqrt{2(1 - \cos(A,B))} = \sqrt{2\mathrm{dist}(A,B)} \tag{2.15}$$

즉, 다음과 같은 관계가 성립됩니다.

$$\mathrm{dist}(A,B) = \frac{1}{2} \| A - B \|^2 \tag{2.16}$$

이러한 단위원에서 코사인 거리와 유클리드 거리의 범위는 모두 [0, 2]입니다. 유클리드 거리는 우리가 거리를 정의할 때 가장 일반적으로 사용하는 범용적인 척도입니다. 따라서 코사인 거리와 유클리드 거리가 이차 관계를 가지고 있으니 자연스럽게 삼각부등식을 만족시키지 못하게 됩니다. 구체적으로 설명하면, A와 B, B와 C가 아주 가깝다고 가정하고, 유클리드 거리는 매우 작은 u라고 가정해 봅시다. 이때 A, B, C가 원호상에 있다고 하더라도 하나의 직선상에 근사하기 때문에 A와 C의 유클리드 거리는 $2u$에 가까울 것입니다. 따라서 A와 B, B와 C의 코사인 거리는 $u^2/2$가 됩니다.

A와 C의 코사인 거리는 $2u^2$에 가깝게 되고, 이는 A와 B, B와 C의 코사인 거리의 합보다 큽니다.

인터뷰를 하면서 이런 종류의 기초 증명 문제를 만나면 쉽게 당황할 수 있습니다. 예를 들어, 앞서 물어본 문제에서 '거리'에 대한 정의가 명확하게 기억나지 않을 수도 있습니다. 이때는 먼저 면접관과 최대한 많이 소통하면서 거리의 정의에 대한 논의를 시작하는 방식을 택해야 합니다(면접관은 지원자가 얼마나 많은 지식을 보유하고 있는지를 보는 것보다는 지원자의 소통 능력과 분석 능력을 더 주의 깊게 볼 가능성이 높습니다). 완벽한 답을 주기 위해서는 명확한 논리와 엄격한 사고능력을 갖추고 있어야 합니다. 예를 들어, 동일성과 대칭성의 증명 과정에서 모호한 설명을 하거나 얼버무리듯이 대답하면 안 됩니다. 마지막으로, 삼각부등식의 증명/위증 과정에서 애매모호하게 '제 느낌에는…' 식의 커뮤니케이션 방식을 사용하는 것보다는 적극적으로 분석하고 명확한 논리로 유클리드 거리와의 관계를 설명해야 합니다. 설령 틀리더라도 증명하는 과정 자체에 대한 정답보다는 논리적으로 증명하는 자세가 중요하기 때문에 좋은 점수를 받을 가능성이 높습니다.

필자가 가장 처음 코사인 거리가 삼각부등식에 부합하지 않다는 것을 알게 된 것은 드라마 라벨링에 대해 연구할 때였습니다. 이 과정에서 comedy와 funny, 그리고 funny와 happy의 코사인 거리가 0.3 이하로 매우 가깝지만, comedy와 happy의 코사인 거리가 0.7 이상인 점을 발견했습니다. 이러한 현상은 거리의 정의에 부합하지 못하는 것인데, 이 때문에 이런 문제를 면접 문제로 출제하게 된 것입니다.

머신러닝 영역에서 흔히 '거리'라고 불리는 것 중에 거리 공식을 만족시키지 못하는 것은 비단 코사인 거리뿐만이 아닙니다. KL 거리Kullback-Leibler divergence*도 거리 공식을 만족시키지 못하는데, 상대 엔트로피relative entropy라고도 불립니다. 이 개념은 두 분포 사이의 차이를 계산하는 데 사용되는데, 대칭성과 삼각부등식 모두를 만족하지 못합니다.

---

* 옮긴이 'KL 발산' 혹은 '쿨백-라이블러 발산'이라고 부르며, 직관적으로는 두 확률분포 사이의 거리 같은 느낌을 주기 때문에 이 책에서는 '거리'라고 표기하였습니다. 본문에서 설명하는 것처럼 거리의 정의는 만족하지 못합니다.

CHAPTER 2

7 과적합과 과소적합

상황 설명

모델 평가와 튜닝 과정에서 우리는 '과적합' 혹은 '과소적합' 상황을 자주 만나게 됩니다. 어떻게 하면 '과적합'과 '과소적합' 현상을 효과적으로 인식하고 목적성 있는 모델 조정을 할 수 있을까요? 이 질문이 바로 머신러닝 모델을 개선하는 핵심 질문입니다. 특히, 실무에서 다양한 방법으로 '과적합'과 '과소적합'의 위험을 줄이는 능력은 데이터 과학자가 필히 갖춰야 할 능력입니다.

키워드 **과적합**Over-Fitting / **과소적합**Under-Fitting

질문 1 모델 평가 과정에서 과적합과 과소적합이란 어떤 현상을 뜻하는 것일까요?

난이도 ★

분석·해답

과적합은 모델이 훈련 데이터에 과하게 맞춰진fitting 현상입니다. 이는 평가 지표에 반영되는데, 일반적으로 훈련 세트상에서의 모델 성능은 매우 좋게 나타나지만, 테스트 세트나 새로운 데이터상에서의 성능이 저조합니다. 과소적합은 모델이 훈련이나 예측에서 모두 좋은 성능을 보이지 못하는 현상을 뜻합니다. 그림 2.5는 과적합과 과소적합을 그래프를 통해 설명하고 있습니다.

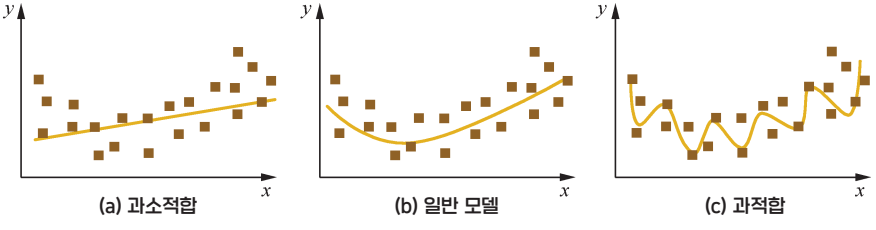


그림 2.5 과소적합과 과적합

과소적합 현상을 보여주는 그림 2.5(a)의 모델은 데이터의 특징을 제대로 잡아내고 있지 못합니다. 반면, 과적합 현상을 보여주는 그림 2.5(c)의 경우는 모델이 지나치게 복잡하여 노이즈 데이터의 특징까지 학습해 모델의 일반화 성능을 저하시킵니다. 이러한 모델은 실제 응용 단계에서 잘못된 예측 결과를 내놓을 가능성이 높습니다.

질문 2 과적합과 과소적합의 위험을 낮출 수 있는 몇 가지 방법에 대해 설명해 주세요.

난이도 ★★

분석·해답

● '과적합' 위험을 낮추는 방법

1 데이터 관점에서 논의를 시작한다면, 더 많은 데이터를 확보하는 것이 과적합 위험을 낮출 수 있는 가장 효과적인 방법입니다 왜냐하면 더 많은 샘플은 모델이 더 많은 유효한 특성을 학습할 수 있도록 하는 동시에 노이즈의 영향을 줄여 주기 때문입니다. 직접적으로 실험에 필요한 데이터를 늘리는 것은 당연히 쉽지 않겠지만, 다른 일정한 규칙을 따라 훈련 데이터를 늘리는 작업은 충분히 가능합니다. 예를 들어, 이미지 분류 문제에서 이미지의 평행 이동, 회전, 수축, 확장 등의 방법으로 데이터 확장을 진행할 수 있습니다. 또는 생성모델generative model을 활용하여 대량의 신규 데이터를 추가하는 방법도 존재합니다.

2 모델의 복잡도를 낮춰 줍니다 데이터가 비교적 적을 경우, 모델이 지나치게 복잡하다면 과적합을 일으킬 확률이 높습니다. 모델의 복잡도를 적당히 낮춰 준다면 모델이 노이즈 데이터에 과도하게 적합되는 것을 방지할 수 있습니다. 예를 들어, 신경망 모델에서 네트워크층이나 뉴런 수를 줄이는 방법을 사용할 수 있습니다. 의사결정 트리의 경우, 나무의 깊이를 줄이고 가지치기를 하는 방법을 사용할 수 있습니다.

3 정규화를 사용합니다 모델의 파라미터에 정규화 항을 추가합니다. 예를 들어, 가중치의 크기를 손실함수에 추가합니다. L2 정규화를 예로 든다면 다음과 같이 나타낼 수 있습니다.

$$C = C_0 + \frac{\lambda}{2n} \cdot \sum_i w_i^2 \tag{2.18}$$

이렇게 한다면 원래의 목적함수 C_0를 최적화하는 동시에 가중치가 너무 커져 과적합 위험이 커지는 것을 어느 정도 제어할 수 있습니다.

4 앙상블 학습 방법을 사용합니다 앙상블 학습은 다수의 모델을 합치는 것인데, 단일 모델의 과적합 위험을 낮춰줄 수 있습니다. 예를 들면, 배깅bagging과 같은 방법이 있습니다.

'과소적합' 위험을 낮추는 방법

1 새로운 특성을 추가합니다 특성이 부족하거나 특성과 샘플 레이블의 상관성이 약할 경우 모델이 과소적합을 일으킬 가능성이 큽니다. 일반적으로 '상하 텍스트 특성', 'ID류 특성', '조합 특성' 등* 새로운 특성을 발굴한다면 더 좋은 효과를 얻을 수 있습니다. 딥러닝이 주류가 되는 추세 속에서 많은 딥러닝 모델이 자동으로 이러한 피처 엔지니어링을 완성해 주는 기능을 더했습니다. 인수분해 머신factorization machine, 그래디언트 부스팅 의사결정 트리gradient boosting decision tree, Deep-crossing 등이 모두 이러한 방법에 속합니다.

2 모델의 복잡도를 증가시킵니다 간단한 모델은 학습 능력이 비교적 떨어지는데, 모델의 복잡도를 올리는 방법을 통해 더 강한 적합 능력을 더해 줄 수 있습니다. 예를 들어, 선형모델에서 고차원의 항을 더한다거나 신경망 모델에서 네트워크층 수나 뉴런 개수를 늘리는 방법들이 있습니다.

3 정규화 계수를 줄입니다 정규화는 과적합을 방지하는 데 사용되는데, 모델이 과소적합 현상을 보인다면 목적성 있게 정규화 계수를 줄여 줘야 합니다.

---

* 옮긴이 상하 텍스트 특성은 상하 문맥을 고려한 특성을 뜻하며, ID류 특성이란 고유 키로 자주 사용되는 ID와 같은 고유 특성을 뜻합니다. 정식 용어는 아니기 때문에 개괄적인 뜻만 이해하고 넘어가면 됩니다.

CHAPTER 5

4 클러스터링 알고리즘 평가

상황 설명

사람은 매우 뛰어난 귀납적 사고능력을 갖췄습니다. 즉, 파편화된 사실 혹은 데이터에서 보편적인 규칙을 찾아 논리적인 결론에 도달하는 작업을 잘합니다. 사용자가 동영상을 보는 행위를 예로 들면, 많은 직관적인 귀납 방식이 존재할 수 있습니다. 예를 들어, 선호 콘텐츠 관점에서 보면 만화, 드라마, 판타지 영화 등으로 나눌 수 있고, 자주 사용하는 기기 관점에서 보면 노트북, 핸드폰, 태블릿 PC 등으로 나눌 수 있습니다. 또한, 사용 시간대 관점에서 보면 저녁, 오후, 매일, 주말마다 보는 사용자로 나눌 수 있을 것입니다. 모든 사용자를 효과적으로 분류할 수 있다면 사용자를 더 깊이 이해하고 더 적합한 콘텐츠를 추천하는 데 중요한 역할을 할 것입니다. 하지만 이러한 문제를 머신러닝으로 처리하기 위해서는 관측 데이터의 라벨 혹은 분류 정보가 없기 때문에 알고리즘 모델을 통해 데이터 내에 존재하는 구조와 패턴을 찾아야 합니다.

키워드 데이터 군집Data Clustering / 클러스터링 알고리즘 평가 지표 Evaluation Metrics for Clustering

질문 외부 라벨(정답) 데이터가 없다고 가정한다면 어떻게 두 클러스터링 알고리즘을 비교할 수 있을까요?

난이도 ★★★

분석·해답

상황 설명 중에서 묘사한 예제는 전통적인 클러스터링 문제입니다. 여기서 확인할 수 있는 것은 데이터의 클러스터링은 실제 목적에 따라 다르고, 동시에 데이터의 특성 척도와 데이터 유사도 평가 방법에 영향을 받는다는 것입니다. 지도학습과 비교

해 보면, 비지도학습은 일반적으로 라벨링된 데이터가 없고, 모델, 알고리즘의 설계가 최종 출력과 모델의 성능에 직접적인 영향을 미칩니다. 서로 다른 클러스터링 알고리즘의 성능을 비교하기 위해서 우리는 먼저 자주 보이는 데이터 군집의 특징을 알아야 합니다.

- **중심에 의해 정의되는 데이터 군집** 이러한 데이터 세트는 구형 분포spherical distribution를 따르는 경향이 있습니다. 일반적으로 중심은 무게중심center of mass으로 정의되는데, 즉 데이터 군집의 모든 샘플 포인트의 평균값입니다. 세트 내의 데이터에서 중심까지의 거리가 다른 군집 중심 대비 짧습니다.
- **밀도에 의해 정의되는 데이터 군집** 이러한 데이터 세트는 주변의 데이터와 명확히 다른 밀도 혹은 희소한 패턴을 보입니다. 데이터 군집이 불규칙하거나 서로 감겨 있는 패턴을 보이고, 노이즈 데이터와 특이점이 있을 경우 밀도에 기반한 군집 정의를 사용합니다.
- **연결에 의해 정의되는 데이터 군집** 이러한 데이터 세트는 데이터 포인트 사이에 연결 관계가 있어 모든 데이터 군집을 구조 그래프로 나타낼 수 있습니다. 이러한 정의는 불규칙한 형태나 서로 감겨 있는 데이터 군집 형태에 효과적입니다.
- **개념에 의해 정의되는 데이터 군집** 이러한 데이터 집합에서는 모든 데이터 포인트가 모종의 공통 특성을 갖습니다.

데이터와 요구의 다양성 때문에 모든 데이터 유형, 데이터 군집 혹은 응용 환경에 통용되는 알고리즘은 존재하지 않습니다. 따라서 각각의 상황에 따라 다른 평가 방법 혹은 척도 기준이 필요합니다. 예를 들어, K평균 클러스터링은 오차제곱합sum of squares error으로 평가할 수 있지만, 밀도에 기반한 데이터 군집은 구형이 아닐 수 있기 때문에 오차제곱합을 사용할 수 없을 수도 있습니다. 많은 상황에서 클러스터링 알고리즘 결과의 좋고 나쁨은 주관적 해석에 의존합니다. 그렇다 하더라도 클러스터링 알고리즘의 평가는 필요하고 클러스터링 분석에서 매우 중요한 부분 중 하나입니다.

클러스터링 평가의 주요 임무는 데이터 세트에서 클러스터링에 대한 타당성을 고려하는 것과 클러스터링 결과의 질quality에 대한 계산을 하는 것입니다. 이러한 틀에서 해당 과정은 세 가지 작은 임무task로 분리될 수 있습니다.

1 클러스터링 경향성 측정 이 단계는 데이터 분포 중에 비-임의성 군집 구조가 존재하는지를 테스트하는 것입니다. 만약 데이터가 기본적으로 랜덤이라면, 클러스터링 결과는 아무런 의미가 없을 것입니다. 우리는 클러스터링 오차가 클러스터링 개수가 늘어남에 따라 단조 변화하는지를 관찰합니다. 만약 데이터가 기본적으로 랜덤이라면 비-임의성 군집 구조가 존재하지 않을 것이고, 따라서 클러스터링 오차는 클러스터링 수가 늘어남에 따라 큰 변화의 폭을 보이지 않을 것이기에 데이터의 실제 군집에 대응하는 적합한 K를 찾지 못할 것입니다.

그 외에도 홉킨스 통계Hopkins Statistic를 사용하여 공간상 데이터의 랜덤성을 판단할 수 있습니다.[7] 먼저 모든 데이터에서 랜덤으로 n개를 찾아 $p_1, p_2, \ldots, p_n$와 같이 나타냅니다. 각 포인트 p_i에 대해 샘플 공간에서 가장 가까운 곳에 위치한 포인트와의 거리 x_i를 계산하여 얻은 거리 벡터를 $x_1, x_2, \ldots, x_n$로 나타냅니다. 그런 다음, 샘플이 취할 수 있는 값 범위 내에서 랜덤으로 n개의 포인트를 생성하고 $q_1, q_2, \ldots, q_n$로 표기합니다. 랜덤으로 생성된 각 포인트에 대해 가장 가까운 곳에 있는 샘플 포인트를 찾고 이들 사이의 거리를 계산하여 $y_1, y_2, \ldots, y_n$를 얻습니다. 홉킨스 통계량 H는 다음과 같이 나타낼 수 있습니다.

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i} \tag{5.15}$$

만약 샘플이 랜덤 분포에 가깝다면 $\sum_{i=1}^{n} x_i$와 $\sum_{i=1}^{n} y_i$의 값은 비교적 비슷할 것입니다. 즉, H의 값이 0.5에 가까워집니다. 만약 클러스터링 경향성이 명확하다면, 랜덤으로 생성된 샘플 포인트 거리는 실제 데이터 포인트의 거리보다 클 것입니다.

즉, $\sum_{i=1}^{n} y_i \gg \sum_{i=1}^{n} x_i$이 되고, H의 값은 1에 가까울 것입니다.

2 데이터 군집 수 판단 클러스터링 경향성을 측정한 후, 우리는 실제 데이터 분포와 가장 유사한 군집 수를 찾고 이에 기반하여 클러스터링 결과의 질을 판단해야 합니다. 데이터 군집 수를 결정하는 방법은 많은데, 예를 들면 엘보우 방법Elbow method과 Gap Statistic 방법 등이 있습니다. 한 가지 설명해야 할 것은 평가에 사용하는 최

적의 데이터 군집 수는 프로그램이 출력한 군집 수와 다를 수 있다는 것입니다. 예를 들어, 어떤 클러스터링 알고리즘은 자동으로 데이터의 군집 수를 결정하는데, 이는 우리가 다른 방법을 통해 결정한 최적의 데이터 군집 수와 다를 수 있습니다.

3 클러스터링 품질 측정 사전에 설정한 군집 개수가 같더라도 클러스터링 알고리즘에 따라 서로 다른 결과를 얻습니다. 어떤 클러스터링 결과가 더 품질이 좋은지 어떻게 측정할 수 있을까요? 비지도 상황에서 우리는 군집이 분리된 상황과 군집이 모여 있는 상황을 통해 클러스터링 효과를 평가할 수 있습니다. 평가 지표를 정의하는 것은 지원자의 실질적인 문제 해결 능력과 분석 능력을 보여줄 수 있습니다. 사실 측정 지표에는 매우 많은 종류가 있는데, 자주 보이는 측정 지표를 소개하겠습니다. 더 많은 지표에 대해서는 관련 문헌을 읽어 보기 바랍니다[8].

- **실루엣 계수**Silhouette Coefficient 하나의 포인트 P가 주어졌을 때 해당 포인트의 실루엣 계수는 다음과 같이 정의됩니다.

$$s(p) = \frac{b(p) - a(p)}{\max\{a(p), b(p)\}} \tag{5.16}$$

  여기서 $a(p)$는 포인트 p와 동일한 군집의 다른 포인트 p' 사이의 평균 거리입니다. $b(p)$는 포인트 p와 다른 군집의 포인트 사이의 최소 평균 거리(만약 다른 군집이 n개가 있다고 한다면, 포인트 p와 가장 가까운 클러스터 내의 포인트와 해당 포인트와의 평균 거리만을 계산합니다)입니다. $a(p)$는 p가 속한 군집에서 데이터가 밀집한 정도를 반영하고, $b(p)$는 해당 군집과 기타 인접한 군집과 떨어진 정도를 반영합니다. 따라서 $b(p)$가 클수록 $a(p)$는 작고, 대응하는 군집 품질이 좋습니다. 따라서 우리는 모든 포인트에 대응하는 실루엣 계수 $s(p)$의 평균값으로 클러스터링 결과의 품질을 측정할 수 있습니다.

- **평균제곱근 표준편차**Root Mean Square Standard Deviation, RMSSTD 클러스터링 결과의 동질성, 즉 밀집 정도를 평가하는 데 사용되며, 다음과 같은 식으로 정의됩니다.

$$RMSSTD = \left\{ \frac{\sum_i \sum_{x \in C_i} \| x - c_i \|^2}{P \sum_i (n_i - 1)} \right\}^{\frac{1}{2}} \tag{5.17}$$

여기서 C_i는 i번째 군집을 나타내고, c_i는 해당 군집의 중심을 나타냅니다. $x \in C_i$는 i번째 군집에 속한 하나의 샘플 포인트를 나타내고, n_i는 i번째 군집의 샘플 수를 나타내며, P는 샘플 포인트에 대응하는 벡터 차원수입니다. 분모는 포인트의 차원수 P에 대한 패널티를 부여하는데, 차원이 높을수록 전체 거리 제곱 값은 커집니다. $\sum_i (n_i - 1) = n - NC$에서 n은 총 샘플 포인트의 수를 나타내고, NC는 군집 수를 나타냅니다. 일반적으로 $NC \ll n$이고, 따라서 $\sum_i (n_i - 1)$의 값은 총 포인트 수에 근접한 상수가 됩니다. 종합하면, RMSSTD는 정규화가 반영된 표준편차라고 볼 수 있습니다.

- **R스퀘어**R-Square 군집 사이의 차이 정도를 측정할 수 있으며, 다음 식으로 정의됩니다.

$$RS = \frac{\sum_{x \in D} \| x - c \|^2 - \sum_i \sum_{x \in C_i} \| x - c_i \|^2}{\sum_{x \in D} \| x - c \|^2} \tag{5.18}$$

여기서 D는 모든 데이터 세트를, c는 데이터 세트 D의 중심점을 나타냅니다. 따라서 $\sum_{x \in D} \| x - c \|^2$은 데이터 세트 D를 단일한 군집으로 봤을 때의 오차제곱합입니다. 위에서 설명한 RMSSTD의 정의와 유사한데, $\sum_i \sum_{x \in C_i} \| x - c_i \|^2$은 데이터 세트를 클러스터링한 후의 오차제곱합입니다. 따라서 R스퀘어는 클러스터링 이후의 결과와 클러스터링 이전의 결과 사이의 비교이며, 대응하는 오차제곱합 지표의 개선 정도를 나타냅니다.

- **개선된 Hubert Γ 통계량** 데이터 쌍의 불일치성을 이용하여 군집의 차이를 평가할 수 있습니다. 개선된 Hubert Γ 통계량은 다음과 같이 정의됩니다.

$$\Gamma = \frac{2}{n(n-1)} \sum_{x \in D} \sum_{y \in D} d(x, y) d_{x \in C_i, y \in C_j}(c_i, c_j) \tag{5.19}$$

식 5.19에서 $d(x, y)$는 데이터 포인트 x에서 y 사이의 거리를 나타내고, $d_{x \in C_i, y \in C_j}(c_i, c_j)$는 데이터 포인트 x가 속한 군집 중심 c_i와 데이터 포인트 y가 속한

군집 중심 c_j 사이의 거리입니다. $\frac{n(n-1)}{2}$는 모든 (x, y)가 일치하는 개수를 나타내고, 따라서 해당 지표는 각 포인트 쌍의 합에 대해 정규화 처리를 한 것과 같습니다. 이상적인 상황에서는 각 포인트 쌍 (x, y)에 대해 $d(x, y)$가 작을수록 대응하는 $d_{x\in C_i, y\in C_j}(c_i, c_j)$ 역시 작아야 합니다(특히, 이들이 같은 군집에 속할 때 $d_{x\in C_i, y\in C_j}(c_i, c_j) = 0$가 됩니다). 반대로, $d(x, y)$가 커질수록 $d_{x\in C_i, y\in C_j}(c_i, c_j)$의 값도 커져야 하고, 따라서 Γ값이 커지는 것은 군집의 결과와 샘플의 원래 거리가 일치하며, 즉 군집 품질이 높다는 것을 설명합니다.

이 외에도 더 합리적으로 서로 다른 클러스터링 알고리즘의 성능을 평가하기 위해 인위적으로 서로 다른 유형의 데이터 세트를 만드는 방법도 필요합니다. 자주 보이는 예제들은 그림 5.10~14에 설명되어 있습니다.

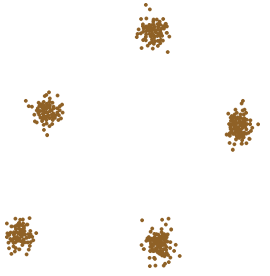


그림 5.10 군집 클래스가 늘어남에 따라 군집 오차가 단조 변화하는지 살펴본다

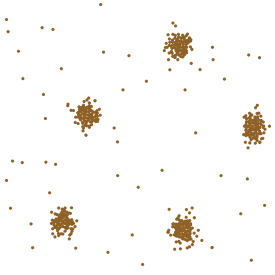


그림 5.11 실제 군집 결과에 대한 군집 오차의 영향을 살펴본다

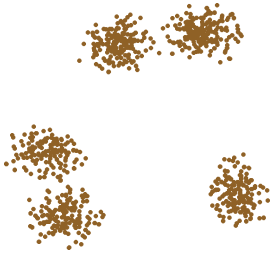


그림 5.12 이웃 데이터 군집의 클러스터링 정확성을 살펴본다

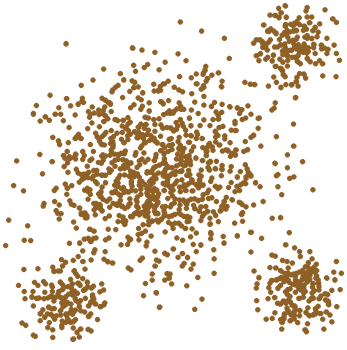


그림 5.13 데이터 밀도에 비교적 큰 차이가 있는 데이터 군집의 클러스터링 효과를 살펴본다

그림 5.14 샘플 수가 비교적 큰 차이를 가진 데이터 군집의 클러스터링 효과를 살펴본다

CHAPTER 9

2 딥러닝의 활성화 함수

상황 설명

선형모델은 머신러닝 영역에서 가장 기초적이고 가장 중요한 도구입니다. 로지스틱 회귀와 선형회귀를 예로 들면, 클로즈드 폼closed-form이나 컨벡스 최적화convex optimization를 사용해 효율적이고 신뢰성 있게 데이터를 적합fitting할 수 있습니다. 그러나 우리는 실전에서 종종 선형으로 분리가 불가능하기 때문에(예를 들면, XOR 함수) 비선형변환을 통해 데이터의 분포에 대해 매핑mapping해야 할 때가 있습니다. 딥러닝 알고리즘에서, 우리는 각 층을 선형변환한 후 하나의 비선형 활성화 함수를 더해 다층 네트워크가 단층 선형함수와 동일하게 되는 것을 피할 수 있고, 그로 인해 더 강력한 학습과 적합 능력을 갖출 수있습니다.

키워드 미적분Calculus / 딥러닝Deep Learning / 활성화 함수Activation Function

질문 1 자주 사용하는 활성화 함수와 해당 활성화 함수의 도함수를 작성해 주세요.

난이도 ★

분석·해답

시그모이드Sigmoid 활성화 함수의 식은 다음과 같습니다.

$$f(z)=\frac{1}{1+\exp(-z)} \tag{9.3}$$

이를 미분하면 다음과 같습니다.

$$f'(z)=f(z)(1-f(z)) \tag{9.4}$$

하이퍼볼릭 탄젠트Hyperbolic Tangent, 이하 Tanh 활성화 함수의 식은 다음과 같습니다.

$$f(z)=\tanh(z)=\frac{e^{z}-e^{-z}}{e^{z}+e^{-z}} \tag{9.5}$$

이를 미분하면 다음과 같습니다.

$$f'(z) = 1 - (f(z))^2 \tag{9.6}$$

렐루Rectified Linear Unit, 이하 ReLU 활성화 함수의 식은 다음과 같습니다.

$$f(z) = \max(0, z) \tag{9.7}$$

이를 미분하면 다음과 같습니다.

$$f'(z) = \begin{cases} 1, z > 0 \\ 0, z \leqslant 0 \end{cases} \tag{9.8}$$

왜 시그모이드와 Tanh 활성화 함수는 그래디언트 소실 현상을 일으킬까요?

난이도 ★★

분석·해답

시그모이드 활성화 함수의 그래프가 그림 9.7에 나와 있습니다. 이 함수는 입력 z를 구간(0, 1)에 매핑시키는데, z가 큰 수일 때 $f(z)$는 1에 가깝게 됩니다. 반대로, z가 매우 작을 경우 $f(z)$는 0에 가깝게 됩니다. 이를 미분하면 $f'(z) = f(z)(1 - f(z))$을 얻는데, z가 아주 크거나 아주 작을 때 모두 0에 근사하게 되어 그래디언트 소실* 현상이 발생합니다.

Tanh 활성화 함수의 곡선은 그림 9.8에 나와 있습니다. z가 큰 수일 때 $f(z)$는 1에 가깝고, z가 작을 때 $f(z)$는 −1에 가깝게 됩니다. 이를 미분하면 $f'(z) = 1 - (f(z))^2$을 얻을 수 있는데, z가 매우 크거나 매우 작을 때 모두 0에 근사하게 됩니다. 따라서 시그모이드 함수와 똑같이 '그래디언트 소실' 현상이 발생합니다. 사실, Tanh 활성화 함수는 시그모이드 함수를 평행 이동한 것과 같습니다.

$$\tanh(x) = 2\text{sigmoid}(2x) - 1 \tag{9.9}$$

* 옮긴이 gradient vanishing은 책에 따라 경사 소실, 기울기 소실 등으로 번역되고 있습니다.

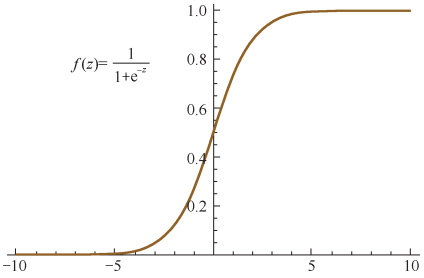


그림 9.7 시그모이드 활성화 함수

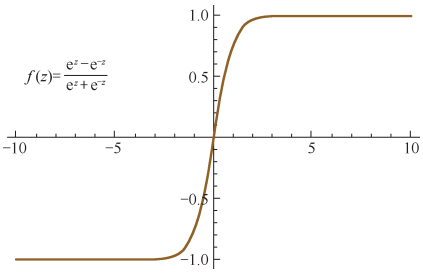


그림 9.8 Tanh 활성화 함수

질문 3 ReLU 계열의 활성화 함수는 시그모이드, Tanh 활성화 함수와 비교했을 때 어떤 장점이 있나요? 이들의 한계는 무엇이며, 어떤 개선 방안들이 있나요?

난이도 ★★★

분석·해답

장점

❶ 계산적 관점에서 보면 시그모이드와 Tanh 활성화 함수는 평균적으로 복잡도가 높으나, ReLU 활성화 함수는 하나의 임계치만 있으면 활성화 값을 얻을 수 있습니다.

❷ ReLU의 비포화성non-saturation은 그래디언트 소실 문제를 효과적으로 해결하고, 상대적으로 넓은 활성화 경계를 제공합니다.

❸ ReLU의 단측면 억제가 네트워크의 희소 표현 능력을 제공합니다.

한계점

ReLU의 한계점은 훈련 과정 중에 뉴런들이 '죽는die' 문제가 발생한다는 것입니다. 이는 함수 $f(z) = \max(0, z)$의 음의 그래디언트가 ReLU 유닛을 경과할 때 0이 되어 버려 이후에 어떤 데이터로도 활성화되지 않기 때문인데, 즉 해당 뉴런을 지나는 그래디언트는 영원히 0이 되어 다른 데이터에 영향을 미치지 않는 것을 의미합니다. 실제 훈련 과정에서 학습률learning rate을 크게 설정하면 일정 비율의 뉴런이 '사망'

해 파라미터 그래디언트를 업데이트할 수 없어 전체 훈련 과정이 실패하는 경우가 발생합니다.

이런 문제를 해결하기 위해 ReLU의 변종인 Leaky ReLU$_{\text{LReLU}}$를 사용하기도 합니다.

$$f(z)=\begin{cases} z, & z>0 \\ az, & z\leqslant 0 \end{cases} \tag{9.10}$$

ReLU와 LReLU의 함수 곡선 비교는 그림 9.9에 나와 있습니다. LReLU와 ReLU의 차이점은 $z<0$일 때 값이 0이 되지 않고 기울기가 a인 선형함수가 됩니다. 일반적으로 a는 아주 작은 정규 수Normal Number이고, 이런 식으로 단측면 억제를 구현하는 동시에 음의 그래디언트 정보를 모두 버리지 않고 부분적으로 유지할 수 있게 됩니다. 하지만 a 값 선택은 문제의 난이도를 증가시켜 많은 경험이나 훈련 횟수에 의지해 적당한 파라미터를 선택해야 한다는 단점이 존재합니다.

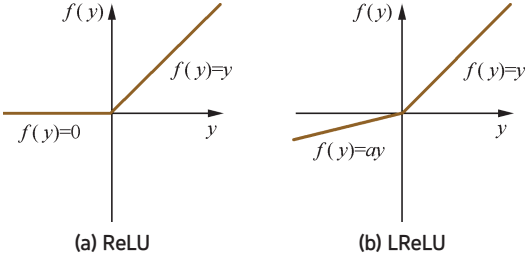


그림 9.9 함수 곡선

이러한 단점을 보완한 파라메트릭 ReLU$_{\text{Parametric ReLU, 이하 PReLU}}$ 활성화 함수도 존재합니다. PReLU와 LReLU의 주요 차이점은 기울기 a를 네트워크 중에서 학습 가능한 파라미터로 설정하고 오차역전파$_{\text{backpropagation}}$ 훈련을 진행해 다른 파라미터를 포함하는 네트워크층과 함께 최적화 작업을 하는 부분입니다. LReLU의 또 다른 변종은 '랜덤화' 메커니즘을 더한 활성화 함수입니다. 훈련 과정에서 기울기 a를 어떤 모종의 분포를 만족하는 랜덤 샘플로 설정하고, 테스트할 때 다시 고정합니다. Random ReLU$_{\text{RReLU}}$는 일정 정도 정규화 작용을 합니다. ReLU 계열 활성화 함수에 대해 더 자세한 내용이나 실험 성능을 비교해 보고 싶은 독자들은 참고문헌을 살펴보길 바랍니다[18].

CHAPTER 12

4 편향과 분산

상황 설명

우리는 과적합, 과소적합을 사용하여 정성적으로 모델이 특정 문제를 얼마나 잘 해결했는지를 묘사합니다. 정량적인 시각에서 보면 모델의 편향bias과 분산variance으로 모델의 성능을 나타낼 수 있습니다. 앙상블 학습은 '신기'하게도 약한 분류기의 성능을 향상시킵니다. 이번 절에서는 편향과 분산의 시각에서 이러한 현상을 설명할 것입니다.

모델의 편향과 분산이란 무엇일까요? 부스팅과 배깅 방법은 편형-분산과 어떤 관계가 있을까요? 이 문제의 해답을 통해 어떻게 편향과 분산 두 지표에 기반해 모델의 최적화와 개선을 진행할 수 있는지에 대해서도 알아볼 것입니다.

키워드 편향Bias / 분산Variance / 리샘플링Resampling / 부스팅Boosting / 배깅Bagging

질문 1 편향과 분산이란 무엇일까요?

난이도 ★★

분석·해답

지도학습 중에서 모델 일반화 오차에 기인하는 요소는 크게 편향과 분산 두 가지가 있습니다. 편향과 분산의 구체적인 정의는 다음과 같습니다.

편향은 크기가 m인 모든 샘플링을 통해 얻은 데이터 세트로 훈련시킨 모델 출력의 평균값과 실제 모델 출력 사이의 편차를 말합니다. 편향은 일반적으로 학습 알고리즘에 대해 잘못된 가설을 설정했을 때 발생하는데, 예를 들어 실제 모델은 어떤 2차 함수인데 모델이 1차 함수라고 가정하는 경우에 편향이 발생합니다. 편향으로 인해 생기는 오차는 일반적으로 훈련오차에 나타나게 됩니다.

분산은 크기가 m인 모든 샘플링을 통해 얻은 데이터 세트로 훈련시킨 모든 모델 출력의 분산을 뜻합니다. 분산은 일반적으로 모델의 복잡도가 훈련 샘플 수 m에 비

해 높을 때 발생하는데, 예를 들면 총 100개의 훈련 샘플이 있는데 모델의 계수가 200 이하의 다항식 함수라고 가정하는 경우입니다. 분산으로 인한 오차는 일반적으로 훈련오차 대비 테스트 오차의 증가에서 나타납니다.

위에서 설명한 정의는 매우 정확하지만 직관적이지 않습니다. 편향과 분산에 대한 이해를 돕는 과녁 예제를 통해 양자 간의 차이와 관련성에 대해 더 자세히 설명하겠습니다. 만약 한 번의 사격이 하나의 샘플에 대한 머신러닝 모델의 예측이라고 가정해 봅시다. 과녁 중심으로 갈수록 예측의 정확도가 높다는 것을 뜻하고, 중심에서 멀어지면 멀어질수록 예측오차가 크다는 것을 뜻합니다. n번의 샘플링을 통해 크기가 m인 n개의 훈련 세트를 얻고, n개의 모델을 훈련시켜 동일한 샘플에 대해 예측을 진행합니다. 즉, n번의 사격을 한 것과 동일한데, 사격 결과는 그림 12.4에 나와 있습니다. 우리가 가장 원했던 결과는 좌측 상단에 보이는 결과일 것입니다. 사격 결과가 정확하면서 집중되어 있습니다. 이는 모델의 편향과 분산이 모두 매우 작다는 것을 뜻합니다. 우측 상단의 과녁은 사격 결과가 과녁 중심에 있긴 하지만 분포가 비교적 퍼져 있습니다. 이는 모델의 편향은 작지만 분산이 큰 것을 뜻합니다. 동일하게, 좌측 하단 과녁은 모델의 분산이 작고 편향이 큰 경우이며, 우측 하단 과녁은 모델의 분산은 크지만 편향 역시 큰 경우를 나타냅니다.

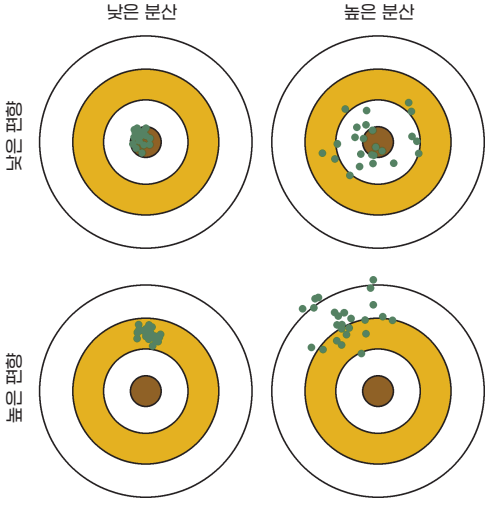


그림 12.4 편향-분산 트레이드오프

질문 2 '편향과 분산 줄이기' 관점에서 부스팅과 배깅의 원리를 설명해 주세요.

난이도 ★★★

분석·해답

이 문제에 대해 간단하게 대답한다면, 배깅이 약한 분류기의 성능을 향상시킬 수 있는 이유는 분산을 낮추기 때문이고, 부스팅이 약한 분류기 성능을 향상시킬 수 있는 이유는 편향을 낮추기 때문이라고 대답할 수 있습니다. 하지만 왜 이렇게 이야기할 수 있는 것일까요?

먼저, 배깅은 'Bootstrap Aggregating'의 약칭입니다. 리샘플링resampling해서 각 샘플에서 훈련된 모델의 평균값을 취합니다.

만약 n개의 랜덤변수가 있고 분산이 σ^2라고 가정하고, 두 변수 사이의 상관관계가 ρ라면 n개 랜덤변수의 평균값 $\frac{\sum X_i}{n}$의 분산은 $\rho*\sigma^2 + (1-\rho)*\sigma^2/n$입니다. 랜덤변수가 완전히 독립적인 상황에서 n개의 랜덤변수의 분산은 σ^2/n이 되는데, 다른 말로 분산이 원래 크기의 $1/n$으로 줄어들었다는 것을 뜻합니다.

모델 자체에 대한 관점에서 이 문제를 이해한다면, n개의 독립적이고 상호 연관되어 있지 않은 모델의 예측 결과의 평균을 취해 분산이 원래 단일 모델의 $1/n$이 되게 만듭니다. 이러한 설명이 완벽한 것은 아니지만 원리를 설명하기에는 충분하다고 생각됩니다. 당연한 이야기이지만, 모델 사이에 완전한 독립적 관계란 존재할 수 없습니다. 하지만 최대한 모델의 독립성을 보존하기 위해 많은 배깅 방법에서 여러 개선 방안이 시도되었습니다. 예를 들어, 랜덤 포레스트 알고리즘에서 매번 노드 분할 속성을 선택할 때 랜덤으로 하나의 속성 하위 집합을 선택하지 모든 속성 중에 최적의 속성을 선택하지 않습니다. 이러한 방법은 약한 분류기 사이에 상호 연관성이 강해지는 것을 방지합니다. 훈련 세트에 대한 리샘플링을 통해 약한 분류기 사이에 일정 독립성을 더해 배깅 후 모델의 분산을 낮춥니다.

이번에는 부스팅에 대해 알아봅시다. 여러분들은 부스팅 훈련 과정에 대해 기억하고 있을 것입니다. 하나의 약한 분류기를 훈련시킨 후, 약한 분류기의 오차 혹은 잔차를 계산해 다음 분류기의 인풋input으로 넣습니다. 이 과정 자체가 손실함수를 계속

해서 줄여 모델을 계속해서 '과녁 중심'으로 가까이 갈 수 있도록 만듭니다. 즉, 모델의 편향을 계속해서 줄여 줍니다. 그러나 부스팅 과정은 분산을 눈에 띄게 줄이지는 못합니다. 그 이유는 부스팅의 훈련 과정은 각각 약한 분류기 사이의 강한 상관성을 갖도록 해 독립성이 부족해집니다. 따라서 분산을 줄이는 작용을 하지 못합니다.

일반화 오차, 편향, 분산, 그리고 모델 복잡도에 대한 관계 그래프는 그림 12.5에 설명되어 있습니다. 그래프를 통해 분산과 편향은 상부상조하며, 모순되면서도 통일되는 관계이기 때문에 양자가 완전히 독립적으로 존재할 수 없음을 쉽게 알 수 있습니다. 주어진 학습 문제와 훈련 데이터에 따라 우리는 모델의 복잡도에 대한 합리적인 가설을 세워야 합니다. 만약 모델의 복잡도가 과도하게 낮다면 분산이 작더라도 편향이 높을 것입니다. 만약 모델 복잡도가 지나치게 높다면 편향이 낮더라도 분산이 높을 것입니다. 따라서 종합적으로 편향과 분산을 고려해 적합한 복잡도를 가진 모델로 훈련을 진행해야 합니다.

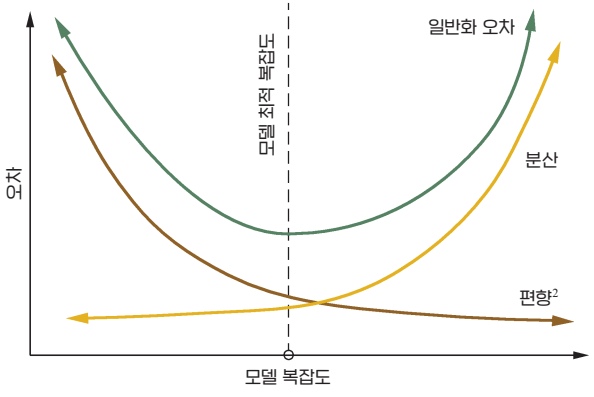


그림 12.5 **일반화 오차, 편향, 분산 그리고 모델 복잡도 사이의 상관 관계**

CHAPTER 12

6 XGBoost와 GBDT의 차이점, 그리고 연관성

상황 설명

XGBoost는 천티엔치Chen Tianqi 등이 개발한 오픈 소스 머신러닝 프로젝트입니다. 효과적으로 GBDT 알고리즘에 대한 많은 개선을 이루어 냈고, 캐글 경진대회 및 기타 머신러닝 경진대회에서 좋은 성적을 거뒀습니다. 우리는 XGBoost 패키지를 사용할 때 XGBoost 내부에 구현된 코드와 원리에 대해서도 잘 알고 있어야 할 것입니다. 그래야만 실무 환경에서 알고리즘을 응용할 때 모델을 자유자재로 튜닝해 목적에 맞게 사용할 수 있을 것입니다.

키워드 XGBoost, GBDTGradient Boosting Decision Tree / 의사결정 트리Decision Tree

질문 XGBoost와 GBDT의 차이점, 그리고 연관성에는 어떤 것들이 있나요?

난이도 ★★★

분석·해답

기존의 GBDT 알고리즘은 경험 손실함수의 음의 경사에 기반해 새로운 의사결정 트리 구조를 만들고, 의사결정 트리가 구성된 후에야 가지치기했습니다. 그러나 XGBoost는 의사결정 트리 구성 단계에서 정규화 항 식 12.2를 더했습니다.

$$L_t = \sum_i l(y_i, F_{t-1}(x_i) + f_t(x_i)) + \Omega(f_t) \tag{12.2}$$

여기서 $F_{t-1}(x_i)$는 현재 가진 $t-1$개 트리의 최적해를 나타냅니다. 트리 구조에 대한 정규항은 다음과 같이 정의됩니다.

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{12.3}$$

여기서 T는 잎 노드 개수이고, w_j는 j번째 잎 노드의 예측값을 나타냅니다. F_{t-1}에서 해당 손실함수에 대해 2차 테일러 전개를 진행하면 다음 식을 유도할 수 있습니다.

$$L_t \approx \tilde{L}_t = \sum_{j=1}^{T}\left[G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2 \right] + \gamma T \tag{12.4}$$

여기서 T는 의사결정 트리 f_t의 잎 노드 개수이며, $G_j = \sum_{i \in I_j} \nabla_{F_{t-1}} l(y_i, F_{t-1}(x_i))$, $H_j = \sum_{j \in I_j} \nabla^2_{F_{t-1}} l(y_i, F_{t-1}(x_i))$이 되고, I_j는 잎 노드 j에 속하는 총 샘플의 인덱스 결합을 뜻합니다.

의사결정 트리의 구조를 이미 알고 있다고 가정한다면, w_j에 대한 손실함수의 도함수를 0으로 설정함으로써 손실함수를 최소화한 상황에서의 각 잎 노드상의 예측값 식 12.5를 구할 수 있을 것입니다.

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{12.5}$$

그러나 모든 트리 구조 중에서 최적의 트리 구조를 찾는 것은 NP-hard* 문제입니다. 따라서 현실에서는 보통 그리디 방법을 사용하여 차선의 트리 구조를 만듭니다. 핵심 아이디어는 근 노드부터 시작해 매번 하나의 잎 노드에 대해 분할을 진행하고, 가능한 각 분할에 대해 특정한 규칙에 기반을 둬 최적의 분할을 선택하는 것입니다. 서로 다른 의사결정 트리 알고리즘은 다른 규칙을 사용하는데, 예를 들어 ID3 알고리즘은 정보 이득information gain을 사용하고, C4.5 알고리즘은 정보 이득에서 값이(혹은 속성 수가) 비교적 많은 특성을 편향적으로 선택하는 단점을 보완한 정보 이득비information gain ratio를 사용합니다. 또한, CART 알고리즘은 지니 계수Gini index와 제곱오차를 사용하고, XGBoost에서도 특정한 규칙을 사용하여 최적의 분할을 선택하고 있습니다.

예측값을 손실함수 중에 대입함으로써 손실함수의 최솟값을 얻을 수 있습니다.

$$\tilde{L}_t^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{12.6}$$

---

* 옮긴이 NP-난해는 NP에 속하는 모든 판정 문제를 다항 시간 내에 다대일로 환산할 수 있는 문제들의 집합입니다.

분할 전후 손실함수의 차이를 쉽게 계산할 수 있습니다.

$$\text{Gain} = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} - \gamma \tag{12.7}$$

XGBoost는 위 차이를 최소화하는 것을 규칙으로 설정하고 의사결정 트리를 만들어 갑니다. 모든 특성의 취할 수 있는 값을 두루 살펴보고 손실함수 전후 차이가 가장 클 때 대응하는 분할 방식을 찾습니다. 이 외에도 전후 손실함수 차이가 반드시 양수이어야 한다는 제약이 있기 때문에 γ는 일정 정도의 사전 가지치기 효과를 냅니다.

계산법에서 전통적인 GBDT와 다른 것 외에도 XGBoost는 프로그래밍 구현 단계에서도 많은 개선을 진행했습니다. 전반적으로 양자 사이의 차이점과 연관성은 다음과 같이 정리할 수 있겠습니다.

1. GBDT는 머신러닝 알고리즘이고, XGBoost는 해당 알고리즘의 엔지니어링적 구현이다.
2. CART를 기초 분류기로 사용했을 때 XGBoost는 정규화 항을 더해 모델의 복잡도를 조절하고, 이는 과적합을 방지하고 모델의 일반화 능력을 향상시키는 데 도움이 된다.
3. GBDT는 모델 훈련 시에 비용함수cost function의 일차 도함수 정보만을 사용하지만, XGBoost는 비용함수에 대해 2차 테일러 전개를 진행해 1차, 2차 도함수를 동시에 사용한다.
4. 전통적인 GBDT는 CART를 기초 분류기로 설정하지만, XGBoost는 선형분류기와 같은 다양한 종류의 기초 분류기를 지원한다.
5. 전통적인 GBDT는 매번 반복할 때마다 모든 데이터를 사용하지만, XGBoost는 랜덤 포레스트와 비슷한 전략을 사용하여 데이터에 대한 샘플링을 지원한다.
6. 전통적인 GBDT는 결측치에 대한 처리 전략이 없지만, XGBoost는 스스로 결측치 처리 전략을 학습한다.

잠시 쉬어가기...

머신러닝 경진대회 캐글

XGBoost의 인기는 머신러닝 경진대회인 캐글Kaggle과 밀접한 연관이 있습니다. 각 대회에서 두각을 나타내는 성능을 선보여 XGBoost는 가장 유행하는 머신러닝 알고리즘이 될 수 있었습니다. 이 기회를 통해 여러분께 캐글에 대한 이야기를 들려주고 싶습니다.

캐글은 세계적인 머신러닝 경진대회입니다. 구글이 캐글을 인수하면서 캐글의 인지도와 사용자 수는 매우 빠르게 늘어 이미 100만 사용자를 넘어섰습니다. 이미 그 규모와 인지도 측면에서 다른 경진대회와 비교할 수 없는 수준에 도달했습니다.

캐글은 창업자의 소박한 아이디어에서 시작되었습니다. 2010년, 오스트레일리아 재정부에 재직 중이던 앤서니 골드블룸Anthony Goldbloom은 당시 자신의 직무에 대해 실망감을 가지고 있었습니다. 그의 주요 업무는 GDP와 물가상승률, 그리고 실업률 등을 예측하는 일이었는데, 전통적인 경제 데이터 규모가 너무 작고 노이즈가 많아 유의미한 결과를 찾아내기 힘들었습니다. 더 많은 데이터 세트와 문제를 얻기 위해 앤서니는 여가시간을 활용해 캐글을 만들었습니다. 이것이 오늘날 가장 인기 있는 머신러닝 경진대회 플랫폼이 된 캐글의 탄생 배경입니다.

처음 생각과는 다르게, 캐글은 앤서니가 생각했던 것보다 훨씬 많이 성장했습니다. 원래는 그냥 재미있는 문제나 데이터를 제공받아 자신의 연구에 활용할 생각이었는데, 갈수록 많은 사람이 참여하기 시작했고, 그 규모를 감당할 수 없다고 판단한 그는 캐글을 코드, 데이터, 그리고 토론의 활기가 넘쳐나는 생태계를 가진 오픈 플랫폼으로 전환하기로 결심합니다.

현재 캐글은 이미 구글 AI 생태계의 중요한 일환으로 자리 잡았습니다. 앤서니는 자신의 창업 경험을 통해 미래의 창업자들에게 두 가지를 강조했습니다. 첫 번째는 자신이 직접 겪었고 다른 사람도 겪고 있을 거라 생각하는 문제를 해야 한다는 것이고, 두 번째는 해당 문제를 해결하기 위한 열정이 있어야 한다는 점입니다.

하지만 필자는 '시대가 영웅을 만든다'고 말하고 싶습니다. 캐글은 인공지능의 세 번째 물결과 맞물려 성장했기 때문입니다. 부디 여러분이 창업할 땐 앤서니가 말한 두 가지 조언을 명심하시고, 그에게 그랬던 것처럼 좋은 운이 함께하길 바라겠습니다.